Practical issues for Differential Privacy

Presenter: Guoming Wang

20 Jan, 2015

Materials

1. Christine Task
    1. A Practical Beginner’s Guide to Differential Privacy

2. Avrim Blum
    1. An brief tour of Differential Privacy

3. Wang Yuxiang
    1. Differential Privacy: a short tutorial

Outline

- **Differential Privacy**
- **Global Sensitivity**
- **Laplacian Noise**

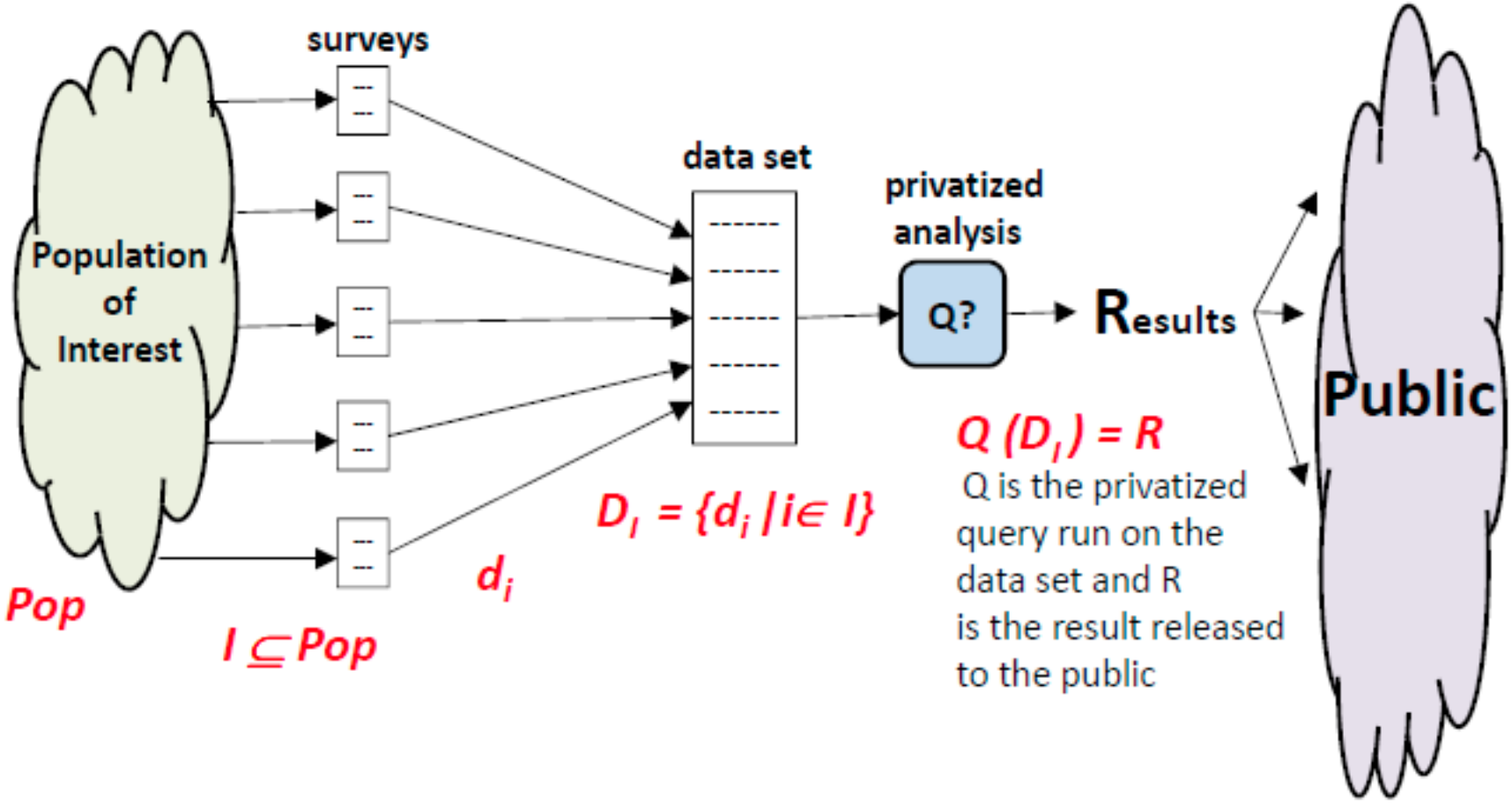
Differential Privacy

Survey questions:

1. How many girls are there in this lab?
2. What is the average age of people in this lab?
3. What is the average GPA of the students in this lab?

If the GPA is his/her sensitive information, people should be guaranteed privacy preserved when he/she is asked the question.

Differential Privacy

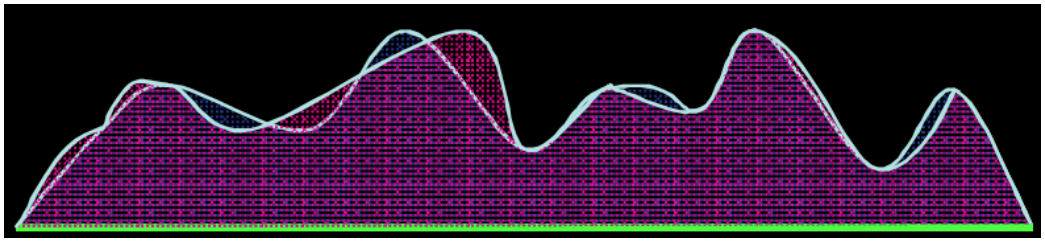


Differential Privacy

Targets:

1. I know that my answer had no impact on the released results.
    1. $Q(D_{(I-me)}) = Q(D_I)$

2. I know that any attacker looking at the published results R could not learn (With any high probability) any new information about me personally.
    1. Prob(secret(me) | R) = Prob(secret(me))



If person i changed his/her input from x_i to ANY OTHER x_i', the relative probabilities of any output do not change by much

Differential Privacy: Definition

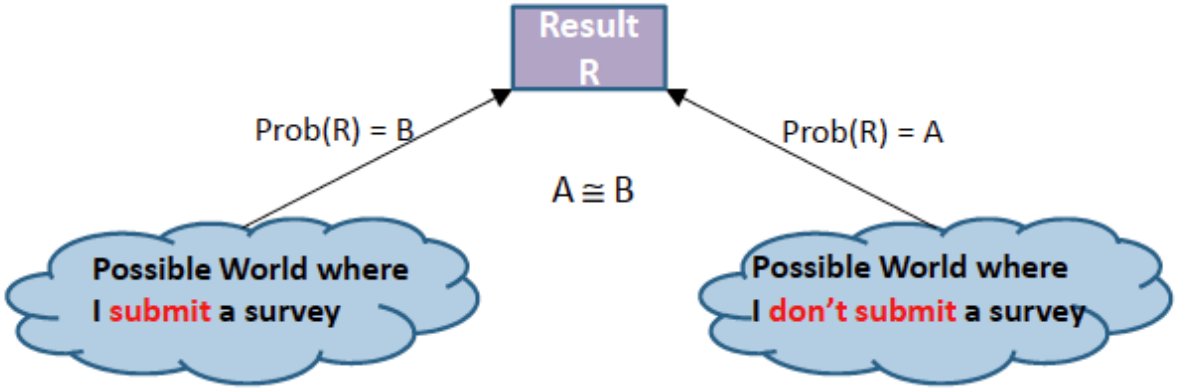
The chance that the noisy released result will be C is nearly the same, whether or not you submit your info.

Definition: ϵ-Differential Privacy

$$\frac{\Pr(M(D) = C)}{\Pr(M(D_{\pm i}) = C)} < e^{\epsilon}$$

For any $|D_{\pm i} - D| \leq 1$ and any $C \in Range(M)$.

Given R, how can anyone guess which possible world it came from?



Global Sensitivity

Given that D1 and D2 are two data sets that differ in exactly ONE PERSON, and F(D) = x is a deterministic, non-privatized function over data set D, which returns a vector X of k real number results.

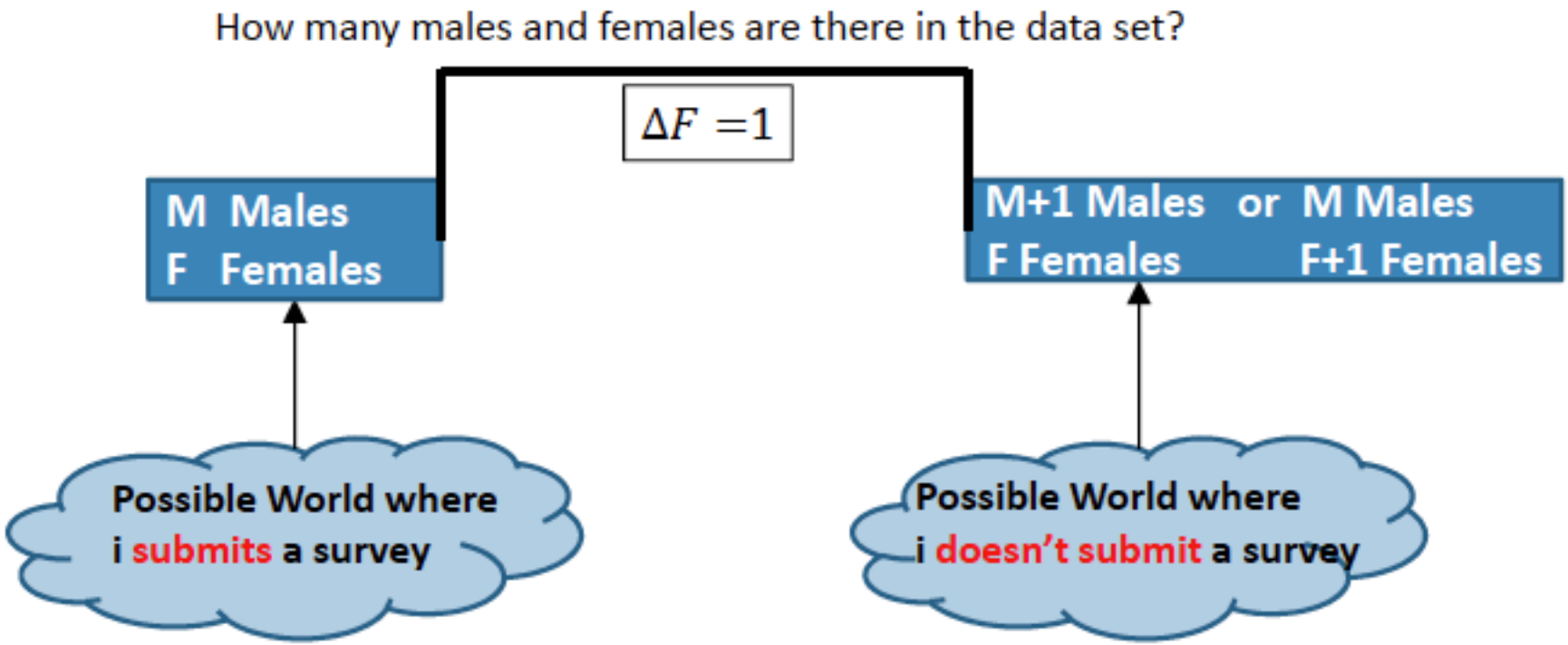
Then the Global Sensitivity of F is:

$$\Delta F = \max_{\{D1,D2\}} ||F(D1) - F(D2)||_{L1}$$

Intuitively, it's the sum of the worst case difference in answers that can be caused by adding or removing someone from a data set.

Global Sensitivity

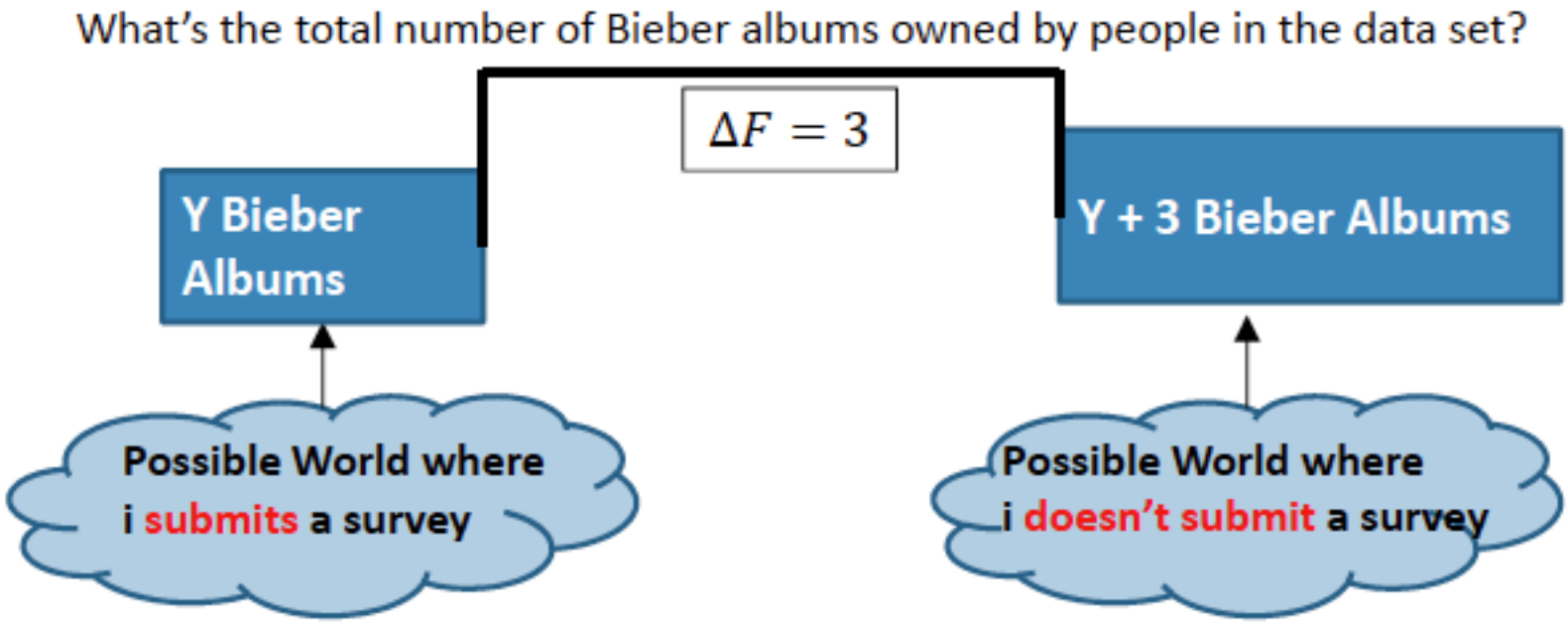
$$\Delta F = \max_{\{D1,D2\}} ||F(D1) - F(D2)||_{L1}$$



Global Sensitivity

$$\Delta F = \max_{\{D1,D2\}} \left|\left|F(D1) - F(D2)\right|\right|_{L1}$$

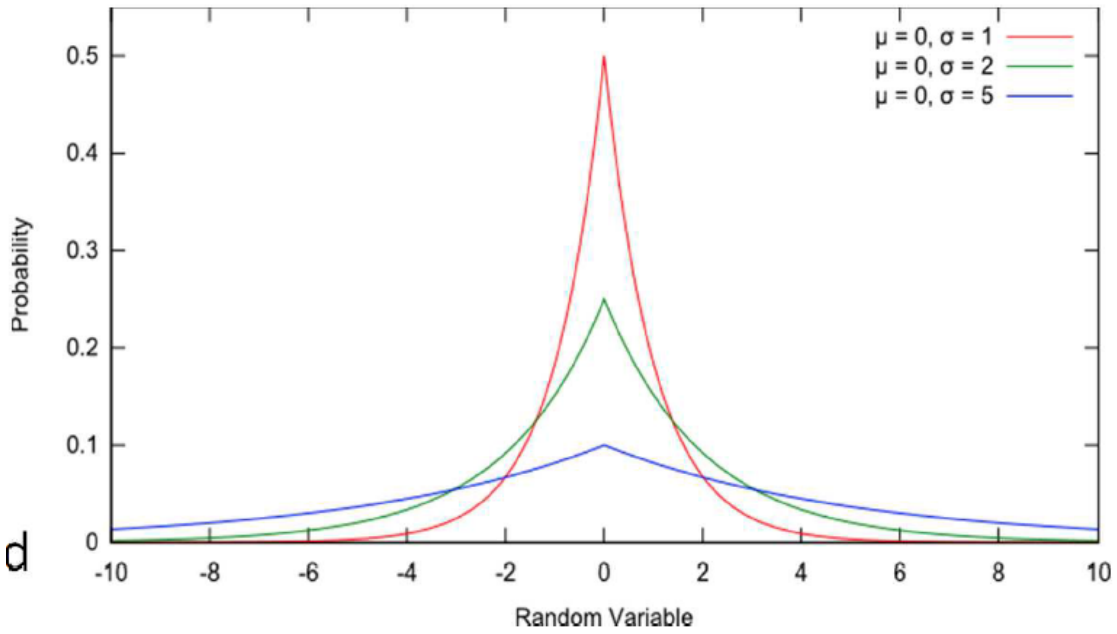
Bieber has 3 albums in 2012.



Laplacian Noise

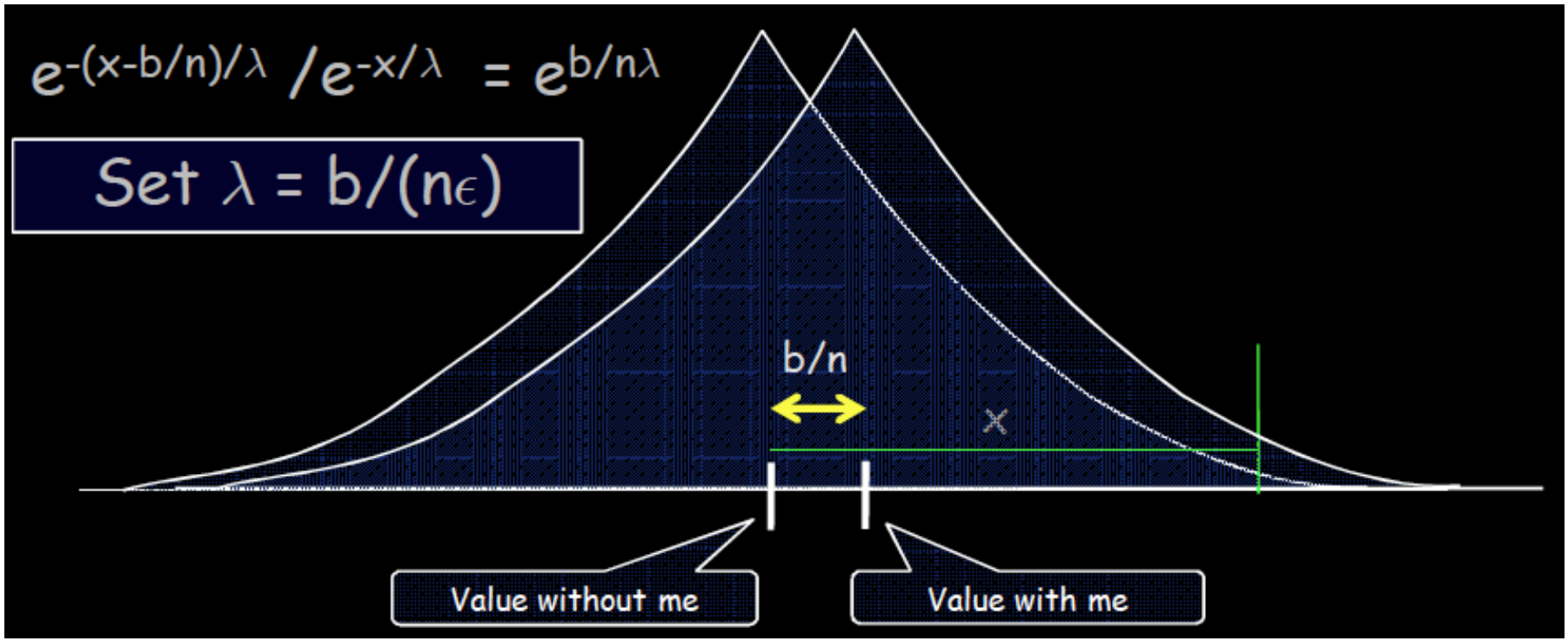
In order for two worst-case neighboring data sets to produce a similar distribution of privatized answers, we need to add noise to span the sensitivity gap.

Adding laplacian Noise is not the only way, but it's easy.



$$Prob(R = x \mid D \text{ is the true world}) = \frac{\varepsilon}{2\Delta F} e^{-\frac{|x - F(D)|\varepsilon}{\Delta F}}$$

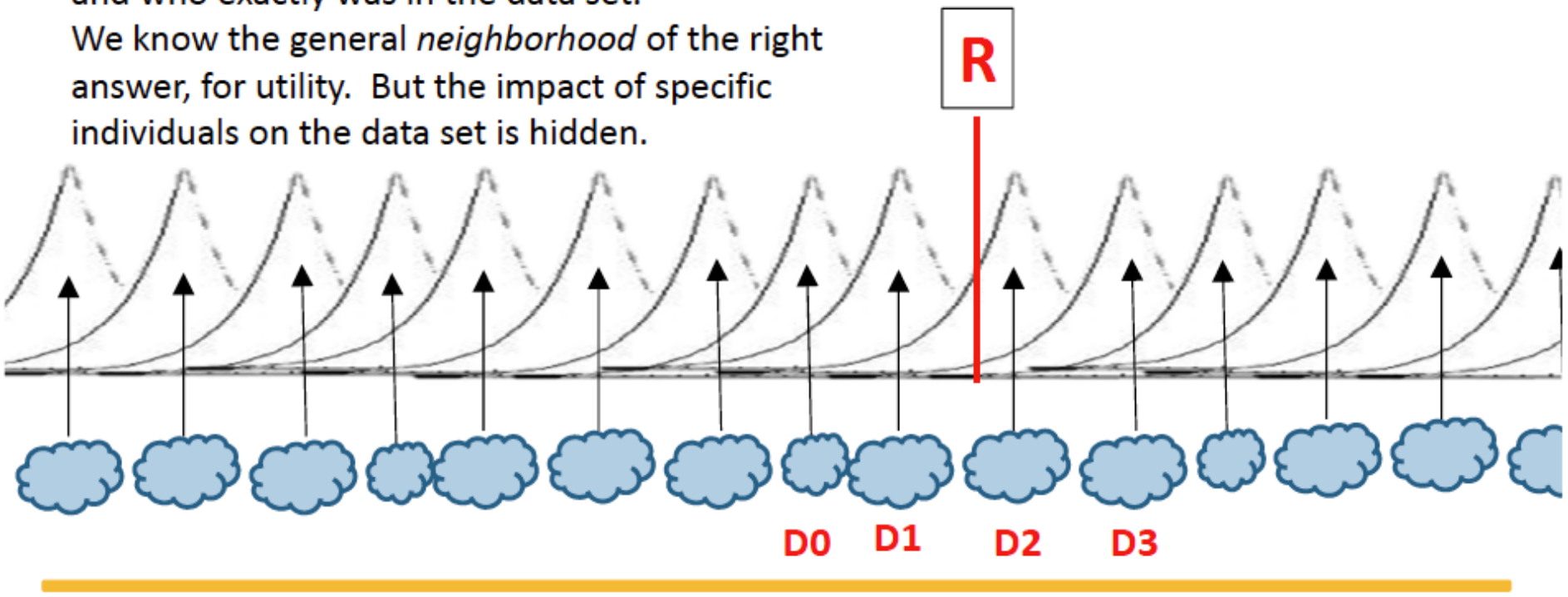
Laplacian Noise



Laplacian Noise

$$Prob(R = x \mid D \text{ is the true world}) = \frac{\varepsilon}{2\Delta F} e^{-\frac{|x - F(D)|\varepsilon}{\Delta F}}$$

Just by looking at the released result R,
it's *very* hard to guess which world it came from
and who exactly was in the data set.
We know the general *neighborhood* of the right answer, for utility. But the impact of specific individuals on the data set is hidden.



Laplacian Noise -- Proof

What we want: $\frac{Prob(R \mid Q(DI))}{Prob(R \mid Q(D_{I\pm i}))} \leq e^{\varepsilon}$

$$\frac{\frac{\varepsilon}{2\Delta F} e^{-\frac{|R-F(D)|\varepsilon}{\Delta F}}}{\frac{\varepsilon}{2\Delta F} e^{-\frac{|R-F(D_{I\pm i})|\varepsilon}{\Delta F}}} \leq e^{\varepsilon} \longrightarrow e^{-\frac{|R-F(D_I)|\varepsilon}{\Delta F}+\frac{|R-F(D_{I\pm i})|\varepsilon}{\Delta F}} \leq e^{\varepsilon}$$

$$\downarrow$$

$$e^{\frac{\varepsilon}{\Delta F} * |F(D_I) - F(D_{I\pm i})|} \leq e^{\varepsilon}$$

Some issues to discuss

How to privatize a series of FIVE overlapping counts across a data set? (ie, “How many people in the data set are female?”, “How many like biber?”, “How many are between age 12-16”, etc)

$$\Delta F = \max_{\{D1,D2\}} ||F(D1) - F(D2)||_{L1}$$

Add laplacian noise calibrated to ΔF = 5, to each count

Some issues to discuss

How to make sure the reuslts still have utility after be adding noise?
(For instance, -1.6 people over 45 like Beiber),

Improve the result self-consistent, or prone off it.
Using common sense, rather than statistic about the true data set.

Counts have low sensitivity, but that's not helpful if you take a noisy count and insert it into a function which is very sensitive to its parameters.

Application

1. Social Network:
    1. C. Task and C. Clifton, “A Guide to Differential Privacy Theory in Social Network Analysis,” in 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012

2. Location Identity:
    1. R. Assam and T Seidl, “A Model for Context-Aware Location Identity Preservation using Differential Privacy” in 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2013

3. Data Mining:
    1. N. Zhang, M. Li and W. Lou, “Distributed Data Mining with Differential Privacy” in IEEE Communications Society subject matter experts for publication in the IEEE ICC 2011 proceedings, 2011

Thank you – Enjoy the rest of your night