L19 -- Differential Privacy
[Jeff Phillips - Utah - Data Mining]

What is Privacy?

Want to release a database D to public so:
 - can compute statistics on D    (think NetFlix Challenge)
 - cannot identify individuals    (John Doe watches adult movies)

Consider hospital patients are surveyed in a study.
 - anonymized by zip code
 - some have cancer

What if senator went to hospital for "routine" treatment, but was only person in zip code

---------- story time --------------
2000:  Massachusettes released all state employees medical records for researchers.
   - wiped ids, but kept zip codes, birthday, gender -- declared safe (by gov)
   - can buy voter data for $20 (has names, birthday, zip code, and birthday)
   + grad student Latanya Sweeney identified governor - sent him his own records.
-----------------------------------

**k-anonymity** can only identify someone up to k other people.  (must be at least k people with same public set of tests in same zip code to release).
  (teacher evaluations work like this)

**l-diversity** each of k people needs to be well-separated.

**t-closeness** distribution of k people needs to look like full distribution (EMD)

What if height was an important quantity?  (Silvester Stallone?)
  Information:  "Sly Stallone is same height as average New Jersey man"

  Independent survey:  "Average New Jersey man is 5' 8" "
    --> Gives away Stallone's height?


---------- story time -------------
Netflix challenge: anonymized user data, but released recommendations
Raters of movies also rate on IMDB (with user id)
By rating similar movies, can identify many people
  (maybe watched adult films on Netflix, not IMDB)
-----------------------------------

---

**Differential Privacy**

Two similar databases D1, D2
  Di = n bits {0,1}
(D1, D2 differ only in one data bit/element)

query q ()

If for all D1, D2 s.t. ||D1 - D2|| = 1
Pr[q(D1) in R]/Pr[q(D2) in R]  <= exp(eps) ~ 1+eps
then
q is a eps-differentially private query

---

Option 1:  (interactive)
Don't publish database D1
 - keep it behind firewall
 - only allow specific queries which are eps-diff-priv
   (any query adds noise to answer)

Recover data with enough queries, but
... requires exponential number of probes to accurately recover data.

---

Option 2:  (non-interactive)
Don't publish database D1 directly
  --> publish D1 + noise (Laplacian) = q(D1)
  then q(D1) can not be more than eps-distinguished from q(D2)
So for any set question R, cannot distinguish D1 from D2

Database is one number (Stallone's height = 68 (inches))
 D1 release 68 + Laplacian noise
  Laplacian(N) = exp(-|N|)
 D2 release 67 + Laplacian noise

Pr[D1 >= 70] ~ exp(-2 * eps)
Pr[D2 >= 70] ~ exp(-3 * eps)

  The range R here is >= 70

Pr[D1 >= 70]/Pr[D2 >= 70] = exp(-2*eps)/exp(-3*eps) = exp(eps) ~ 1+eps
  D1 is eps-differentially private

---

Another view:
If you add your data to database, and adversary cannot detect if you did.
  + because any allowable query will change by at most eps.

---
Example: subset Sum

```
D      = [0 1 0 1 0 1 0 1 0 0 0 1 1 1 0 0 1]   n-bit vector
query  = [1 0 1 1 0 0 1 1 0 1 0 1 1 0 1 0 1]   n-bit vector mask
          0 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0 1 = 5
```

If || D - D' || <= 1,
 then |q(D) - q(D')| <= 1

make differentially private if A(D,q) = q(D) + LapNoise(1/alpha)

$$\frac{Pr(A(D,q) = x)}{Pr(A(D',q) = x)} = \frac{Lap(|x-q(D)|)}{Lap(|x-q(D')|)}$$
$$= \exp(a(|x-q(D)|-|x-q(D')|))$$
$$\leq \exp(a\,|q(D) - q(D')|)$$
$$\leq \exp(a) \sim 1+alpha$$

---

trade-off:   more noise = more private + less informative