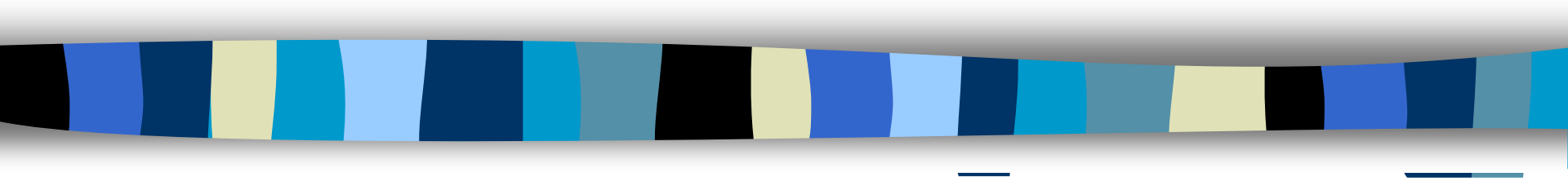
Information Security
CS 526



Topic 21: Data Privacy

What is Privacy?

- Privacy is the protection of an individual's personal information.

- Privacy is the rights and obligations of individuals and organizations with respect to the collection, use, retention, disclosure and disposal of personal information.

- Privacy ≠ Confidentiality

OECD Privacy Principles

- **1. Collection Limitation Principle**
  - There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

- **2. Data Quality Principle**
  - Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

OECD Privacy Principles

- **3. Purpose Specification Principle**
  - The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

- **4. Use Limitation Principle**
  - Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Principle 3 except:
  - a) with the consent of the data subject; or
  - b) by the authority of law.

OECD Privacy Principles

- **5. Security Safeguards Principle**
  - Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.

- **6. Openness Principle**
  - There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

OECD Privacy Principles

- **7. Individual Participation Principle**
  - An individual should have the right:
  - a) to request to know whether or not the data controller has data relating to him;
  - b) to request data relating to him, …
  - c) to be given reasons if a request is denied; and
  - d) to request the data to be rectified, completed or amended.

- **8. Accountability Principle**
  - A data controller should be accountable for complying with measures which give effect to the principles stated above.

Areas of Privacy

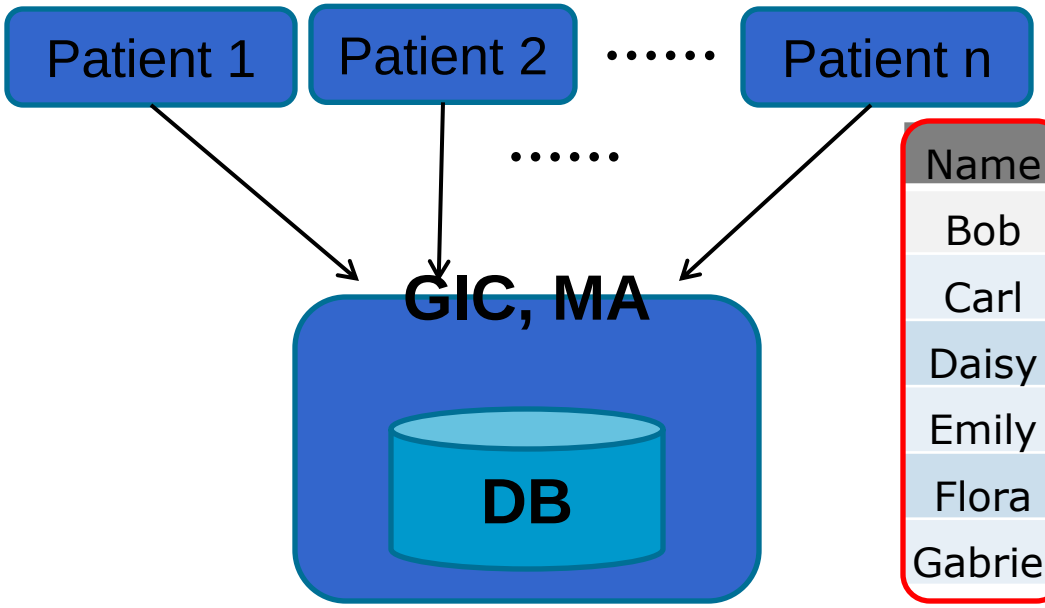
- Anonymity
  - Anonymous communication:
    - e.g., The TOR software to defend against traffic analysis
- Web privacy
  - Understand/control what web sites collect, maintain regarding personal data
- Mobile data privacy, e.g., location privacy
- Privacy-preserving data usage

Privacy Preserving Data Sharing

- The need to sharing data
  - For research purposes
    - E.g., social, medical, technological, etc.
  - Mandated by laws and regulations
    - E.g., census
  - For security/business decision making
    - E.g., network flow data for Internet-scale alert correlation
  - For system testing before deployment
  - …
- However, publishing data may result in privacy violations

GIC Incidence [Sweeny 2002]

- Group Insurance Commissions (GIC, Massachusetts)
  - Collected patient data for ~135,000 state employees.
  - Gave to researchers and sold to industry.
  - Medical record of the former state governor is identified.



Name	Age	Sex	Zip code	Disease
Bob	69	M	47906	Cancer
Carl	65	M	47907	Cancer
Daisy	52	F	47902	Flu
Emily	43	F	46204	Gastritis
Flora	42	F	46208	Hepatitis
Gabriel	47	F	46203	Bronchitis

Re-identification occurs!

AOL Data Release [NYTimes 2006]

- In August 2006, AOL Released search keywords of 650,000 users over a 3-month period.
  - User IDs are replaced by random numbers.
  - 3 days later, pulled the data from public access.

AOL searcher # 4417749

"landscapers in Lilburn, GA"
queries on last name "Arnold"
"homes sold in shadow lake subdivision Gwinnett County, GA"
"num fingers"
"60 single men"
"dog that urinates on everything"

NYT

Thelman Arnold, a 62 year old widow who lives in Liburn GA, has three dogs, frequently searches her friends' medical ailments.

Re-identification occurs!

Netflix Movie Rating Data [Narayanan and Shmatikov 2009]

- Netflix released anonymized movie rating data for its Netflix challenge
  - With date and value of movie ratings
- Knowing 6-8 approximate movie ratings and dates is able to uniquely identify a record with over 90% probability
  - Correlating with a set of 50 users from imdb.com yields two records
- Netflix cancels second phase of the challenge

Re-identification occurs!

Genome-Wide Association Study (GWAS) [Homer et al. 2008]

- A typical study examines thousands of singe-nucleotide polymorphism locations (SNPs) in a given population of patients for statistical links to a disease.
- From aggregated statistics, one individual's genome, and knowledge of SNP frequency in background population, one can infer participation in the study.
  - The frequency of every SNP gives a very noisy signal of participation; combining thousands of such signals give high-confidence prediction

Membership disclosure occurs!

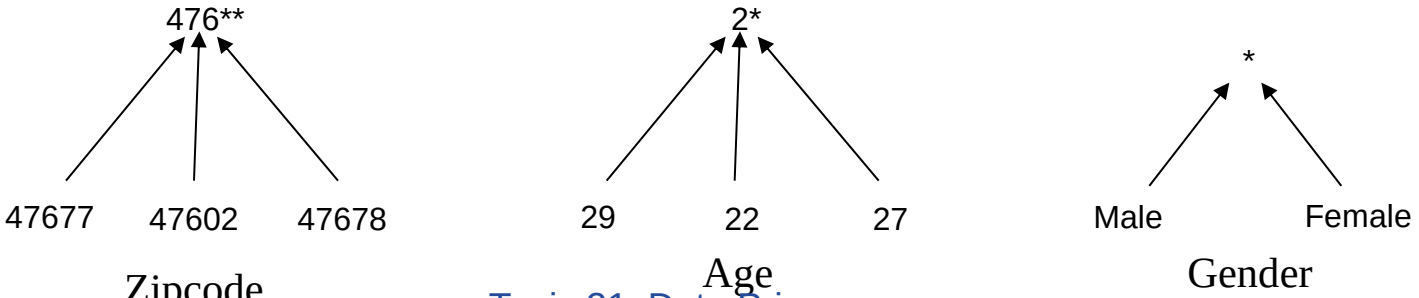
	Study group Avg	Population Avg	Target individual
SNP 1=A	43%	42%	yes
SNP 2=A	11%	11%	no
SNP 3=A	58%	59%	no
SNP 4=A	23%	22%	yes
…			
…			

Need for Data Privacy Research

- Identification Disclosure (GIC, AOL, Netflix)
  - Leaks the subject individual of one record
- Attribute Disclosure
  - leaks more precise information about the attribute values of some individual
- Membership Disclosure (GWAS)
  - leaks an individual's participation is in the dataset
- Research Program: Develop theory and techniques to anonymize data so that they can be beneficially used without privacy violations.
- How to define privacy for anonymized data?
- How to publish data to satisfy privacy while providing utility?

k-Anonymity [Sweeney, Samarati]

- Privacy is "protection from being **brought to the attention of others**."
- k-Anonymity
  - Each record is indistinguishable from ≥ k-1 other records when only "quasi-identifiers" are considered
  - These k records form an equivalence class
- To achieve k-Anonymity, uses
  - Generalization: Replace with less-specific values
  - Suppression: Remove outliers



Example of *k*-Anonymity & Generalization

The Microdata

QID			SA
Zipcode	Age	Gen	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

The Generalized Table

QID			SA
Zipcode	Age	Gen	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

- 3-Anonymous table
  - The adversary knows Alice's QI values (47677, 29, F)
  - The adversary does not know which one of the first 3 records corresponds to Alice's record.

Attacks on *k*-Anonymity

- k-anonymity does not provide privacy if:
  - Sensitive values lack diversity
  - The attacker has background knowledge

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl does not have heart disease

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

l–Diversity: [Machanavajjhala et al. 2006]

- Principle
  - Each equi-class contains at least l well-represented sensitive values
- Instantiation
  - Distinct l-diversity
    - Each equi-class contains *l* distinct sensitive values
  - Entropy l-diversity
    - *entropy(equi-class)*≥*log*$_2$(l)

$$H(X) = \mathrm{E}(I(X)) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

The Skewness Attack on *l*-Diversity

- Two values for the sensitive attribute
  - HIV positive (1%) and HIV negative (99%)
- Highest diversity still has serious privacy risk
  - Consider an equi-class that contains an equal number of positive records and negative records.
- l-diversity does not differentiate:
  - Equi-class 1: 49 positive + 1 negative
  - Equi-class 2: 1 positive + 49 negative

l-diversity does not consider the overall distribution of sensitive values

The Similarity Attack on *l*-Diversity

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

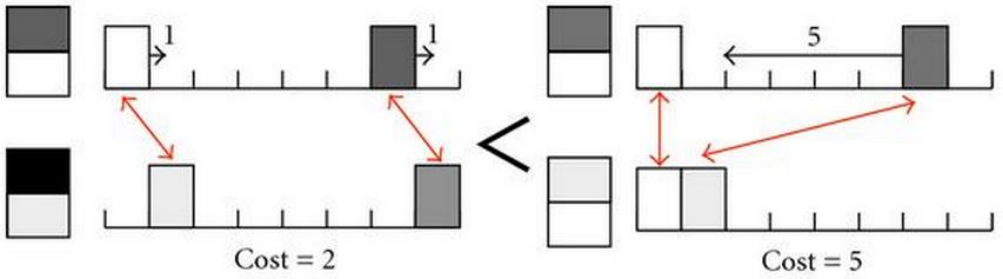
1. Bob's salary is in [20k,40k], which is relative low.
2. Bob has some stomach-related disease.

l-diversity does not consider semantic meanings of sensitive values

t-Closeness

- Principle: Distribution of sensitive attribute value in each equi-class should be close to that of the overall dataset (distance $\leq$ t)
  - Assuming that publishing a completely generalized table is always acceptable
  - We use Earth Mover Distance to capture semantic relationship among sensitive attribute values
- **(n,t)-closeness:** Distribution of sensitive attribute value in each equi-class should be close to that of some natural super-group consisting at least n tuples

N. Li, T. Li, S. Venkatasubramanian: t-Closeness: Privacy Beyond k-Anonymity and *l*-diversity. In ICDE 2007. Journal version in TKDE 2010.



From Syntactical Privacy Notions to Differential Privacy

- Limitation of previous privacy notions:
  - Requires identifying which attributes are quasi-identifier or sensitive, not always possible
  - Difficult to pin down due to background knowledge
  - Syntactic in nature (property of anonymized dataset)
    - Not exhaustive in inference prevented
- Differential Privacy [Dwork et al. 2006]
  - Privacy is not violated is one's information is not included
  - Output does not overly depend on any single tuple

Definition (ε-Differential Privacy)

A randomized algorithm $\mathcal{A}$ satisfies ε-differential privacy, if for any pair of neighboring datasets D and D' and for any $O \subseteq \text{Range}(\mathcal{A})$:

$$e^{-\varepsilon} Pr[\mathcal{A}(D') \in O] \leq Pr[\mathcal{A}(D) \in O] \leq e^{\varepsilon} Pr[\mathcal{A}(D') \in O]$$

Differential Privacy [Dwork et al. 2006]

- Definition: A mechanism *A* satisfies ε-Differential Privacy if and only if
  - for any neighboring datasets D and D'
  - and any possible transcript $t \in$ Range(A),
    $$\Pr[A(D) = t] \leq e^{\epsilon} \Pr[A(D') = t]$$
  - For relational datasets, typically, datasets are said to be neighboring if they differ by a single record.
- Intuition:
  - Privacy is not violated if one's information is not included in the input dataset
  - Output does not overly depend on any single record

The Desire to Have a Semantic Interpretation of DP

- Why needs a semantic interpretation?
  - "*Definition [of DP] equates privacy with the inability to distinguish two close databases. Indistinguishability is a convenient notion to work with; However, it does not directly say what an adversary may do and learn.*" [Dwork et al. 2006: "Calibrating Noises …"].
- What makes a semantic interpretation?
  - A Bayesian approach: An adversary has a prior belief; after interacting with A(D), the adversary has a posterior belief. We want to place some limitation on the posterior belief

Impossibility of Bounding Arbitrary Prior-to-Posterior Belief Change

- Dalenius [in 1977] proposes this as privacy notion: "*Access to a statistical database should not enable one to learn anything about an individual that could not be learned without access*."
  - Similar to the notion of semantic security for encryption
  - Not possible in the context if one wants utility.
- The Terry's height example:
  - *Adversary knows "Terry is two inches shorter than the average Lithuanian woman"*
  - *Published data reveal average height of Lithuanian woman*
  - *Seeing published info enable learning Terry's height*

Different Manifestation of the Impossibility Result

- Dwork & Naor: "*absolute disclosure prevention (while preserving utility at the same time) is impossible because of the arbitrary auxiliary information the adversary may have*".

- Kifer and Machanavajjhala: "*achieving both utility and privacy is impossible without making assumptions about the data*."

- Li et al. (Membership privacy framework): *"without restricting the adversary's prior belief about the dataset distribution, achieving privacy requires publishing essentially the same information for two arbitrary datasets"*

What to do now for Semantic Interpretation of DP?

- Approach 1: Provide posterior-to-posterior bound
  - Identify “ideal worlds” where individuals’ privacy are preserved: the i’th ideal world is to remove the i’th individual’s data
  - Bound difference between ``ideal worlds’’ and the “real world”
- Approach 2: Understand under which condition prior-to-posterior bound can be ensured

An Alternative Formulation

- Adversary is modeled as a decision function f, which after observing a transcript t, makes a decision among C.
- The prob an adversary chooses c after interacting is
  - $Adv^{A(D)}(c) = \sum_t \Pr[A(D) = t] * f(t)[c]$
- *Def:* α-opting-out-simulation: Let D' be the result of an individual opting out from D, then
  - $e^{-\alpha} Adv^{A(D')}(c) \leq Adv^{A(D)}(c) \leq e^{\alpha} Adv^{A(D')}(c)$
- Thm: ε-DP is equivalent to α-opting-out-simulation

DP's Similar-Decision-Regardless-of-Prior Guarantee

- Regardless of external knowledge, an adversary with access to the sanitized database makes similar decisions whether or not one individual's data is included in the original database.

The Personal Data Principle

- Data privacy means giving an individual control over his or her personal data. An individual's privacy is not violated if no personal data about the individual is used. Privacy does not mean that no information about the individual is learned, or no harm is done to an individual; enforcing the latter is infeasible and unreasonable.

An Attempt at Providing Prior-to-Posterior Bound in [Dword et al. 2006]

- A mechanism is said to be **(k, ε)-simulatable** if for **every informed adversary** who already knows all except for k entries in the dataset D, every output, and every predicate f, the change in the adversary's belief on f is multiplicative-bounded by e^{ε}.
- Thm: ε-DP is equivalent to (1,ε)-simulatable.
- Does this mean ε-DP provides prior-to-posterior bound for an arbitrary adversary?
  - Wouldn't that conflict with the impossibility results?

Counter-Arguments

- From [Kifer and Machanavajjhala, 2011]
- "*Additional popularized claims have been made about the privacy guarantees of differential privacy. These include:*
  - *It makes no assumptions about how data are generated.*
  - *It protects an individual's information (even) if an attacker knows about all other individuals in the data.*
  - *It is robust to arbitrary background knowledge.*"

An Example Adapted from [Kifer and Machanavajjhala, 2011]

- Bob or one of his 9 immediate family members may have contracted a highly contagious disease, in which case the entire family would have been infected. An adversary asks the query "how many people at Bob's family address have this disease?"
- What can be learned from an answer produced while satisfying ε-DP?
  - Answer: Adversary's belief change on Bob's disease status may change by something close to $e^{10\varepsilon}$.
- Anything wrong here?

First, The Technical Aspects

1. An adversary's belief about Bob's disease status may change by a factor of $e^{10\epsilon}$ due to data correlation. This is an example that DP cannot bound prior-to-posterior belief change against arbitrary external knowledge.

2. DP's guarantee about posterior-to-posterior bound remains valid.

3. The analysis in [Dwork et al. 2006] is potentially misleading, because it could lead one to think that DP can offer more protection than it actually does.
   - The notion of informed adversary, while appearing strong, is in fact, very limiting.

Still, Does ϵ-DP provide the intended level of privacy protection?

- It is complicated… Consider the following variants.
- Case (a). Bob lives in a dorm building with 9 other unrelated individuals. Either they all have the disease or none. One can query how many individuals at this address have the disease.
- Case (b). The original example: Bob and 9 family members.
- Case (c). Bob and 9 minors for which Bob is the legal guardian.
- Case (d). DP is defined over records, each record corresponds to a single visit; Bob may have 10 visits.

Our Answer

- Case (a). Bob and 9 other unrelated individuals.
  - DP does what it suppose to do based on Personal Data Principle.
- Case (b). The original example: Bob and 9 family members.
  - Difficult to say: on the borderline and not enough information.
- Case (c). Bob and 9 minors
  - Using DP this way is inappropriate, because Bob controls the 9 other records as well, and
- Case (d). DP is defined over records, each record corresponds to a single visit; Bob may have 10 visits.
  - Using DP this way is inappropriate.

1. ε-DP does not prevent inference of personal info when there is correlation.
2. Whether this acceptable or not depends on whether definition of neighboring datasets simulates opting-out correctly.

When is ϵ-DP Good Enough?

- Applying ϵ-DP in a particular setting provides sufficient privacy guarantee when the following three conditions hold:
  - (1) The Personal Data Principle can be applied;
  - (2) All data one individual controls are included in the difference of two neighboring datasets;
    - With (1) and (2), even if some information about an individual is learned because of correlation, one can defend DP.
  - (3) An appropriate ϵ value is used.

How to Choose ε

- From the inventors of DP: "*The choice of ϵ is essentially a social question. We tend to think of ϵ as, say,* 0.01, 0.1, *or in some cases,* ln 2 *or* ln 3".

- Our position.
  - ϵ of between 0.1 and 1 is often acceptable
  - ϵ close to 5 might be applicable in rare cases, but needs careful analysis
  - ϵ above 10 means very little

- Why?

Consult This Table of Change in Belief: p is prior; numbers in table are posterior

ϵ	0.01	0.1	1	5	10
$\gamma = e^{\epsilon}$	1.01	1.11	2.72	148	22026
$p = 0.001$	0.0010	0.0011	0.0027	0.1484	1.0000
$p = 0.01$	0.0101	0.0111	0.0272	0.9933	1.0000
$p = 0.1$	0.1010	0.1105	0.2718	0.9939	1.0000
$p = 0.5$	0.5050	0.5476	0.8161	0.9966	1.0000
$p = 0.75$	0.7525	0.7738	0.9080	0.9983	1.0000
$p = 0.99$	0.9901	0.9910	0.9963	0.9999	1.0000

On Defining Neighbors Incorrectly

- Edge-DP in graph data is inappropriate
  - Typically one individual controls a node and its relationship.
  - ``Attacks'' on graph anonymization typically in the form of node identification.
  - Suppose the goal is to protect edge info, then edge-DP still fails, because of correlation between edges.
- Packet-level privacy for networking data is inappropriate
- Cell-level privacy in matrix data is usually inappropriate
- Google's RAPPOR system is not good enough
  - Data collection views answer to each question separately, the same ε is applied to each question

Apply a Model Learned with DP Arbitrarily.

- There are two steps in Big Data
  - Learning a model from data from individuals in A
  - Apply the model to individuals in B, using some (typically less sensitive) personal info of each individual, one can learn (typically more sensitive) personal info.
    - The sets A and B may overlap
- The notion of DP deals with only the first step.
- Even if a model is learned while satisfying DP, applying it may still result in privacy concern, because it uses each individual's personal info.

The Target Pregnancy Prediction Example

- Target assigns every customer a Guest ID number and stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources.
- Looking at historical buying data for all the ladies who had signed up for Target baby registries in the past, Target's algorithm was able to identify about 25 products that, when analyzed together, allowed Target to assign each shopper a ``pregnancy prediction'' score.
- Target could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Potential Challenge to Personal Data Principle

- Can a group of individuals, none of whom has specifically authorized usage of their personal information, together sue on privacy grounds that aggregate information about the individual is leaked?
  - If so, satisfying DP is not sufficient.
  - Would size of group matter? Probably?

Coming Attractions …

- Role-Based Access Control