18734: Foundations of Privacy

Database Privacy: k-anonymity and de-anonymization attacks

Sruti Bhagavatula
Based on slides by Piotr Mardziel and Anupam Datta
CMU
Fall 2019

Administrative

- Homework 2 deadline postponed
  - Monday, Sept. 30, midnight in PIT or SV, wherever you are enrolled

- Combination recitation/office hours: regular time on Friday, Sept. 27
  - Come get help with AdFisher!

- When submitting, please mark your answers clearly on Gradescope!

In-class Quiz

- Take on Canvas
- Go over answers in class

Last time

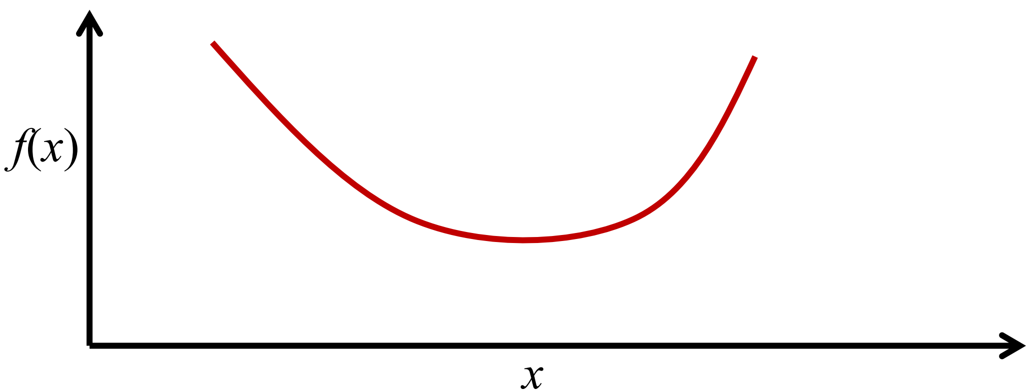
- Score function: Softmax classifier (linear classifier)
  - Maps raw data to class scores
  - Usually parametric
- Loss function (objective function): Cross-entropy loss
  - Measures how well predicted classes agree with ground truth labels
  - How good is our score function?
- Learning
  - Find parameters of score function that minimize loss function

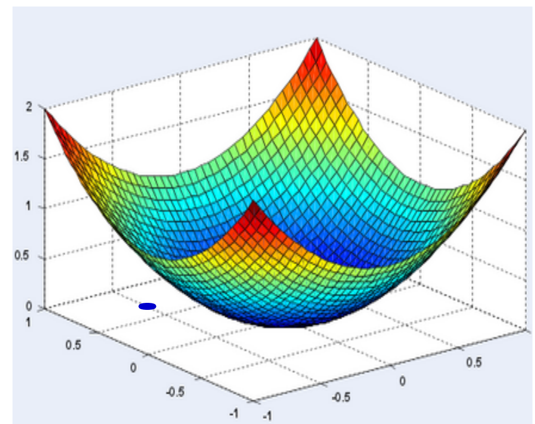
Learning task

- Find parameters of the model that make our loss as small as possible

- There are many different techniques for training models
  - stochastic gradient descent is a popular one
  - scikit-learn provides implementations

The problem of optimization

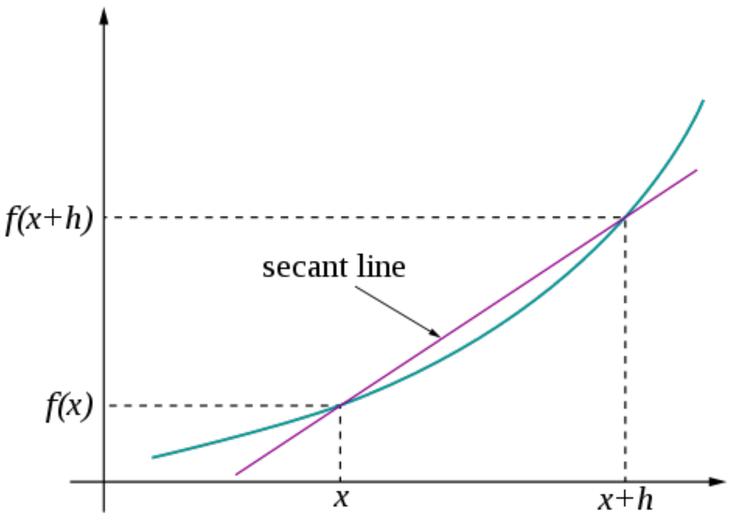


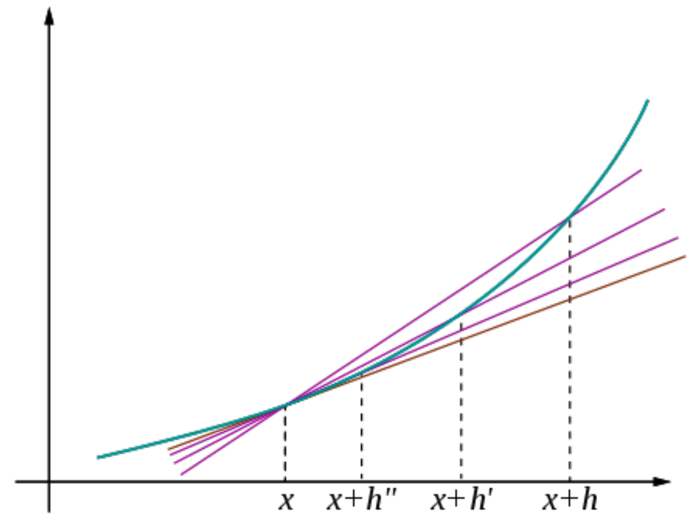


Find the value of ***x*** where **f(*x*)** is minimum

Our setting: ***x*** represents weights $(\text{e.g.}, W, b)$, **f(*x*)** represents loss function (e.g., average cross-entropy)

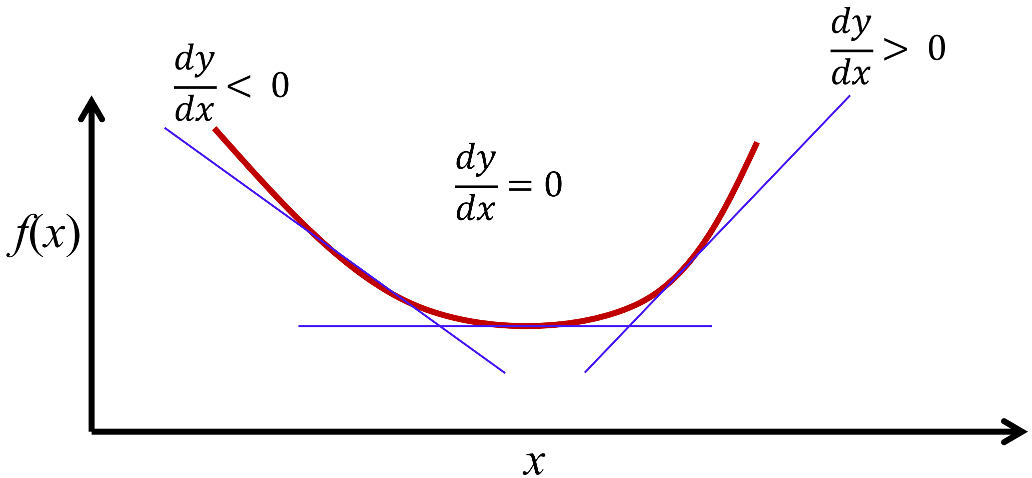
Derivative of a function of single variable





$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

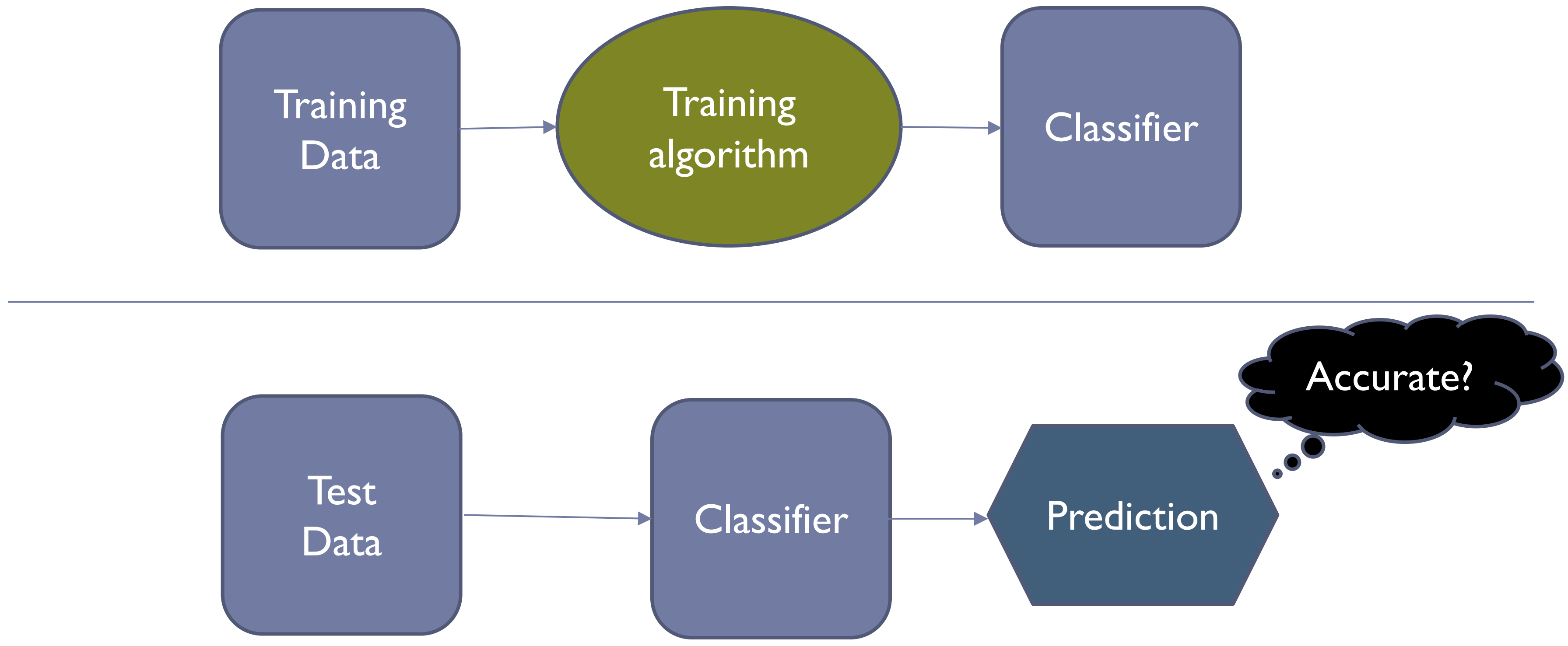
Finding minima



Increase x if derivative negative, decrease if positive
i.e., take step in direction opposite to sign of gradient
(key idea of gradient descent)

Animation courtesy of Christopher Gondek
https://www.youtube.com/watch?v=GCvWD9zIF-s

Classification pipeline



Last time

- Score function: Softmax classifier (linear classifier)
  - Maps raw data to class scores
  - Usually parametric
- Loss function (objective function): Cross-entropy loss
  - Measures how well predicted classes agree with ground truth labels
  - How good is our score function?
- Learning: Gradient Descent (or variants thereof)
  - Find parameters of score function that minimize loss function

Acknowledgment

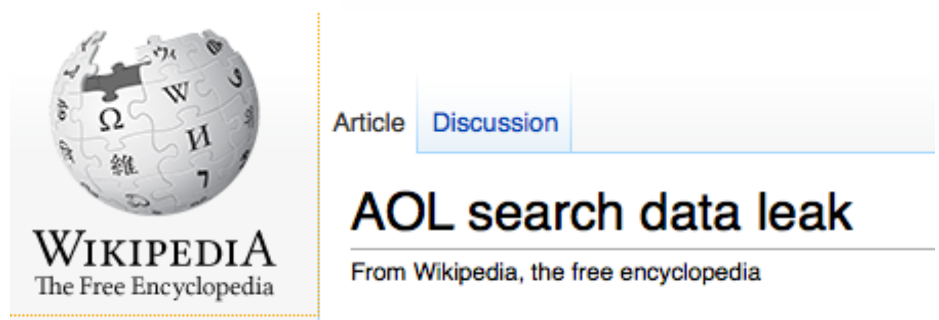
- Based on material from Stanford CS231n
  http://cs231n.github.io/

Today

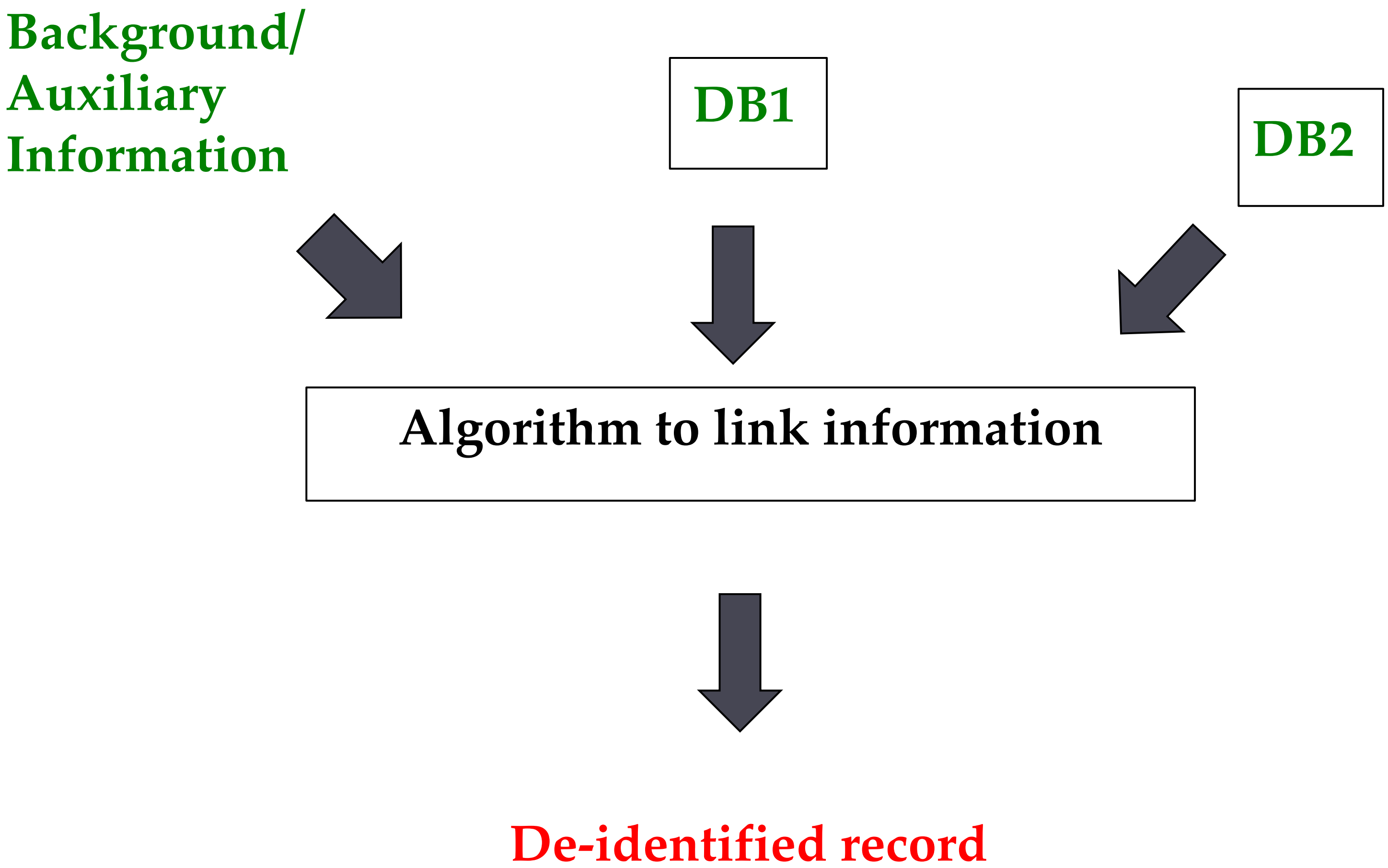
DEANONYMIZING DATASETS

Publicly Released Large Datasets

- Useful for improving recommendation systems, collaborative research
- Contain personal information
- Mechanisms to protect privacy, e.g. anonymization by removing names
- Yet, private information leaked by attacks on anonymization mechanisms



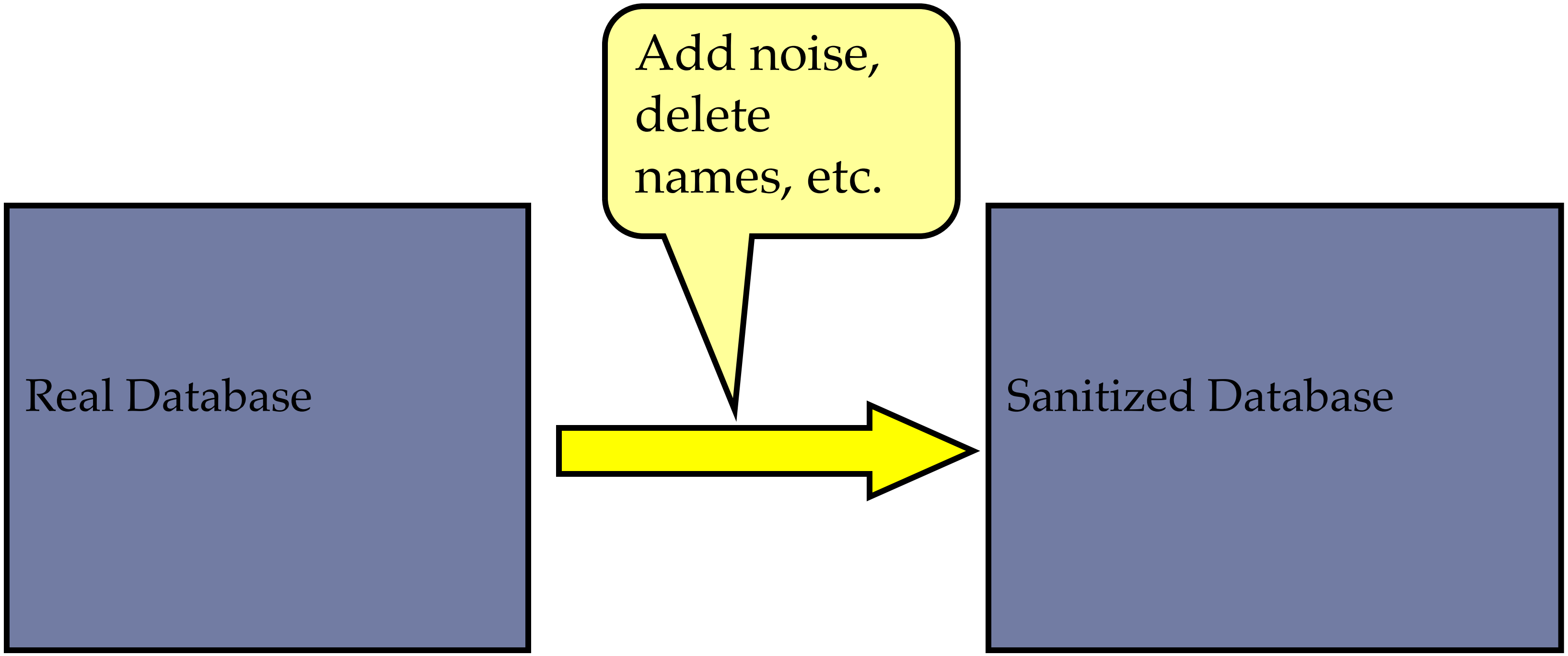
Non-Interactive Linking



Roadmap

- ~~Motivation~~
- Privacy definitions
- Netflix-IMDb attack
- Empirical results
- Conclusion

Sanitization of Databases



Health records

Census data

Protect privacy

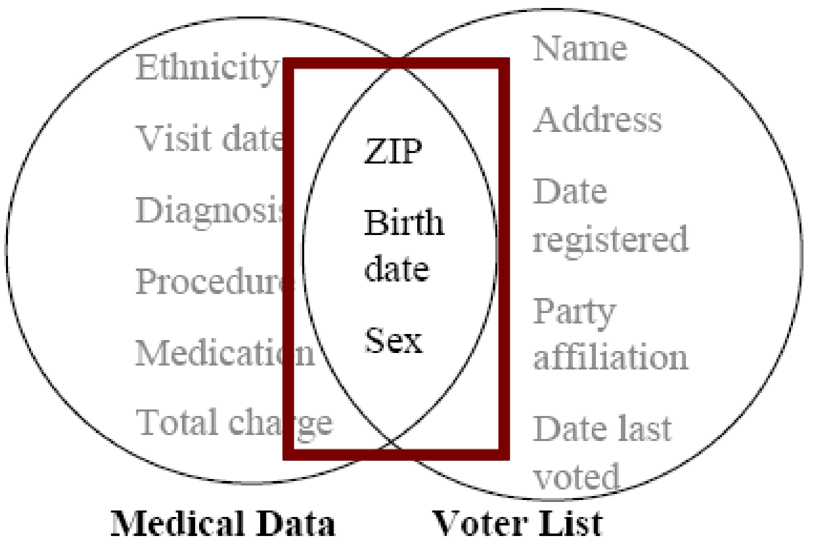
Provide useful information (utility)

Database Privacy

- Releasing sanitized databases
  1. k-anonymity [Samarati 2001; Sweeney 2002]
  2. l-diversity [Machanavajjhala 2007]
  3. t-closeness [Li 2007]
  4. Differential privacy [Dwork et al. 2006] (*future lecture*)

Re-identification by linking

Linking two sets of data on shared attributes may uniquely identify some individuals:



87 % of US population uniquely identifiable by 5-digit ZIP, gender, DOB

K-anonymity

- Quasi-identifier: Set of attributes that can be linked with external data to uniquely identify individuals

- Given a quasi-identifier:
  - Make every record in the table indistinguishable from at least *k-1* other records with respect to quasi-identifiers
  - Linking on quasi-identifiers yields at least *k* records for each possible value of the quasi-identifier

K-anonymity

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Equivalence class

What is the issue with k-anonymity?

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Advantages: Provides some protection: linking on ZIP, age, nationality yields 4 records

Limitations: lack of diversity in sensitive attributes, background knowledge, subsequent releases on the same data set

L-diversity

- Given a k-anonymized table:
  - Ensure that within an equivalence class, there are at least *l* "well-represented" values of the sensitive attribute

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

- *k = 4*
- *l = ?*

What is the issue with l-diversity?

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Limitations:

- Values of the sensitive attribute within one equivalence class may have semantic similarity; can infer some property of the sensitive attribute (i.e., stomach-related disease)
- Could have high *k* and low *l*, resulting in a high occurrence of one value of the sensitive attribute in the equivalence class.

T-closeness

- Given a k-anonymized and l-diverse table:
  - Ensure that the distance between the distribution of each sensitive attribute in the eq. class and the distribution of the attribute value in the whole table is $\leq t$

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

  - *Salary: t = 0.167*
  - *Disease: t = 0.278*

Re-identification Attacks in Practice

Examples:

- Netflix-IMDB
- Movielens attack
- Twitter-Flicker
- Recommendation systems – Amazon, Hunch,..

Goal of De-anonymization: To find information about a record in the released dataset

Roadmap

- ~~Motivation~~
- ~~Privacy definitions~~
- Netflix-IMDb attack
- Empirical results
- Conclusion

Anonymization Mechanism

	Gladiator	Titanic	Heidi
Bob	5	2	1
Alice	3	2.5	2
Charlie	1.5	2	2

Each row corresponds to an individual

Each column corresponds to an attribute, e.g. movie

Delete name identifiers and add noise

	Gladiator	Titanic	Heidi
? r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Anonymized Netflix DB

De-anonymization Attacks Still Possible

- Isolation Attacks
  - Recover individual's record from anonymized database
  - E.g., find user's record in anonymized Netflix movie database
- Information Amplification Attacks
  - Find more information about individual in anonymized database
  - E.g. find ratings for specific movie for user in Netflix database

Netflix-IMDb Empirical Attack [Narayanan et al 2008]

Anonymized Netflix DB

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Publicly available IMDb ratings (noisy)

	Titanic	Heidi
Bob	2	1

Used as auxiliary information

Weighted Scoring Algorithm

Isolation Attack!

r_1	4	1	0

Netflix-IMDb Empirical Attack [Narayanan et al 2008]

Anonymized Netflix DB

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Publicly available IMDb ratings (noisy)

	Titanic	Heidi
Bob	2	1

Used as auxiliary information

What does **auxiliary information** about a record mean?

Weighted Scoring Algorithm

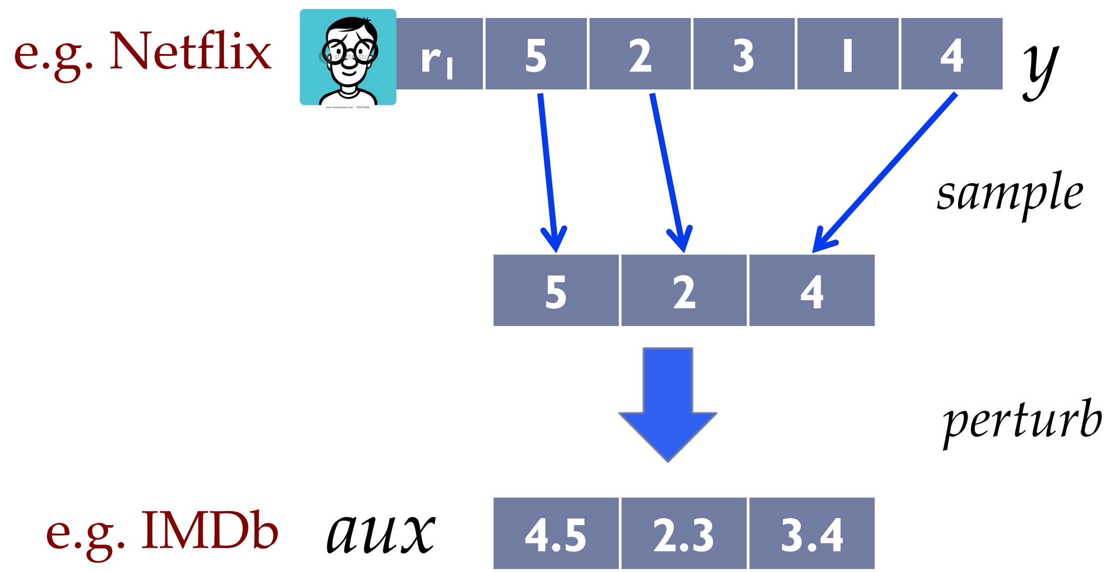
How do you measure similarity of this record with Bob's record? **(Similarity Metric)**

r_1	4	1	0

Definition: Auxiliary Information

Intuition:

- *aux* about *y* should be a subset of record *y*
- *aux* can be noisy



aux captures information available outside normal data release process

Problem Statement

Anonymized database

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Auxiliary information about a record (noisy)

	Titanic	Heidi
Bob	2	1

Attacker uses weighted scoring algorithm to find record

Attacker's goal: Given an anonymized database *D* and auxiliary record *aux(r')*, find $r \in D$ such that r and r' are similar.

Weighted Scoring [Narayanan et al 2008, Frankowski et al 2006]

Intuition: The fewer the number of people who watched a movie, the rarer it is

Weight of an attribute *i*

$$w(i) = \frac{1}{\log|\mathrm{supp}(i)|}$$

$|\mathrm{supp}(i)|$ = no. of non null entries in column i

Use weight as an indicator of rarity

Score gives a weighted average of how closely two people match on every movie, giving higher weight to rare movies

Scoring Methodology

$$\mathrm{Score}(\mathrm{aux}, r') = \sum_{i \in \mathrm{supp(aux)}} w(i)\mathrm{Sim}(\mathrm{aux}_i, r'_i)$$

$|\mathrm{supp}(aux)|$ = m = no. of non null attributes in *aux*

Compute *Score* for every record *r* in anonymized DB to find out which one is closest to target record y.

(aux is derived from *y*)

Weighted Scoring Algorithm [Narayanan et al 2008]

Compute *Score* for every *r* in *D*

$$\text{Score}(\text{aux}, r) = \sum_{i \in \text{supp(aux)}} w(i)\text{Sim}(\text{aux}_i, r_i)$$

w_i	0.63	0.5	0.63

	v_1	v_2	v_3
r_1	5	2	-
r_2	3	1	4
r_3	-	2	4

Score(aux, r_j)
0.52
0.40
0.23

aux

v_1	v_2
4.5	2.3

One of the records *r* in anonymized database is *y*. Which row is it?

Choose a threshold ϕ: **Eccentricity**

If $(\max_{r \in D} Score(aux, r) - max2_{r' \in D} Score(aux, r'))/2 > \phi$

output record with highest score

Else

no match

r_1	5	2	-

Main Result

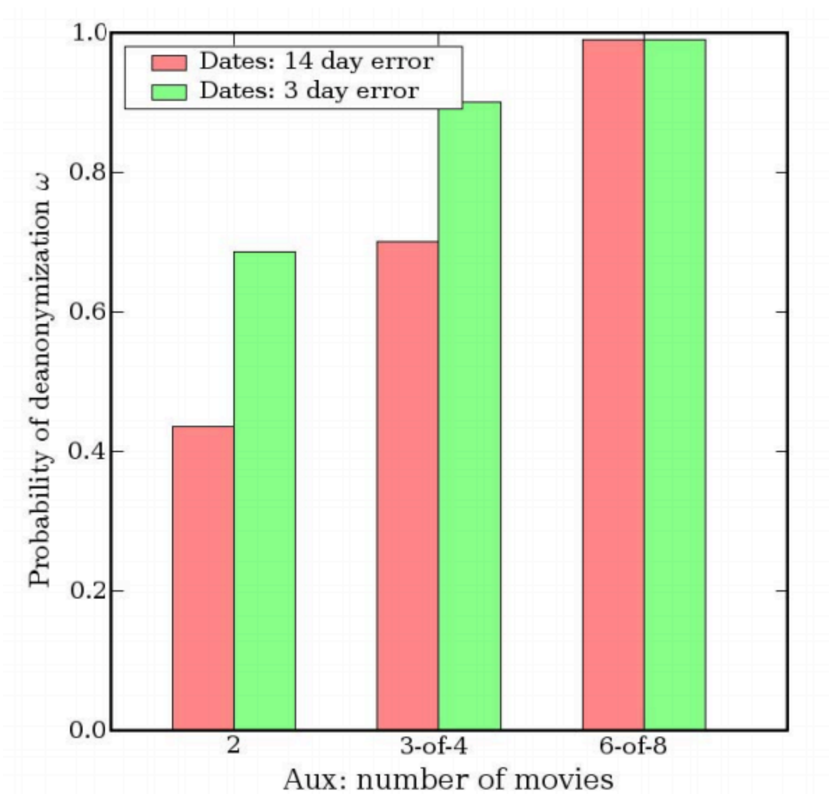
- **Definition.** A database is (θ, ω)-deanonymized w.r.t. auxiliary information aux if there exists an algorithm A which, on inputs D and aux(r) where r is sampled uniformly from D outputs r' such that
$$\Pr[\mathrm{Sim}(r, r') \geq \theta] \geq \omega.$$

- **Theorem.** Let $0 < \epsilon, \delta < 1$ and let D be the database. Let aux consist of at least $m \geq \frac{(\log N - \log \epsilon)}{-\log(1-\delta)}$ randomly selected attributes of target record r, with $\mathrm{Sim}(\mathrm{aux}_i, r_i) \geq 1 - \epsilon \ \forall i \in supp(aux)$. Then D can be $(1 - \epsilon - \delta, 1 - \epsilon)$-deanonymized w.r.t. aux.

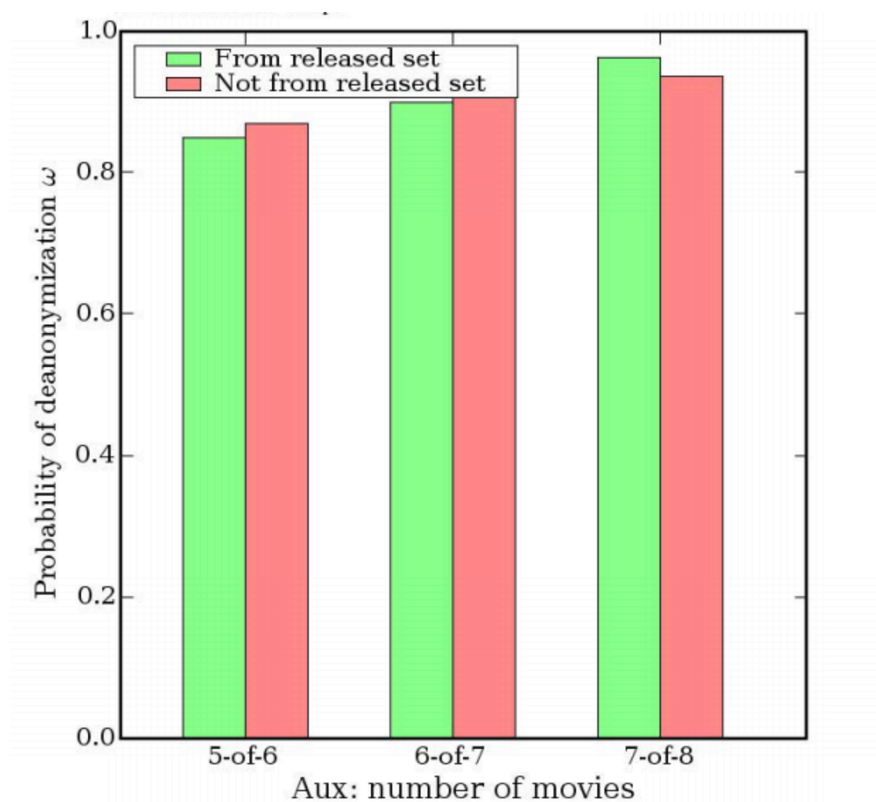
Roadmap

- ~~Motivation~~
- ~~Privacy definitions~~
- ~~Netflix-IMDb attack~~
- Empirical results
- Conclusion

Empirical Results

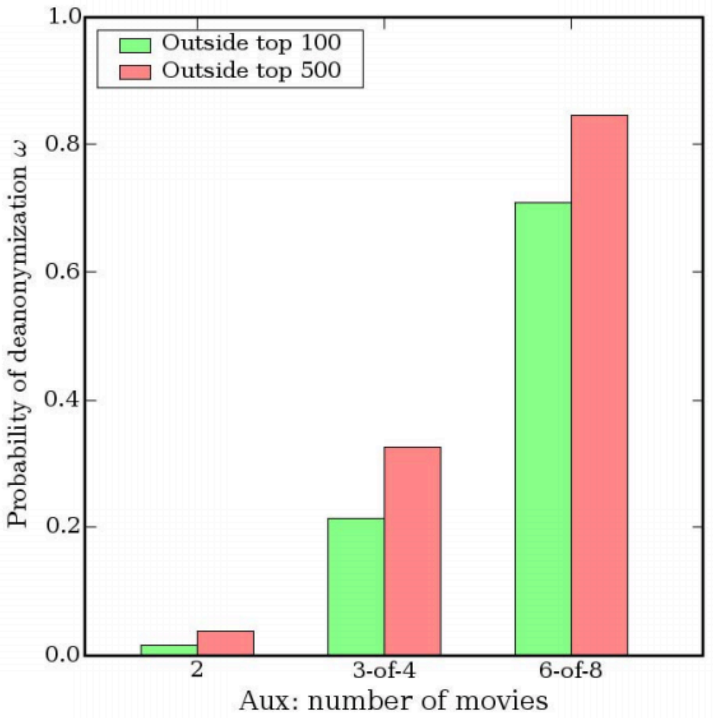


Adversary knows exact ratings and approximate dates.

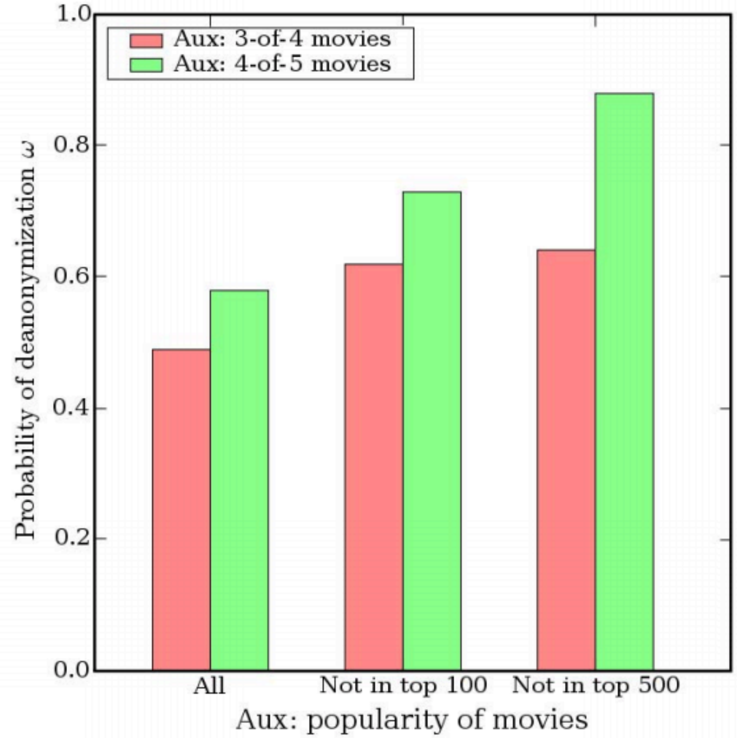


Same parameters as previous graph, but the adversary must also detect when the target record is not in the sample

Empirical Results

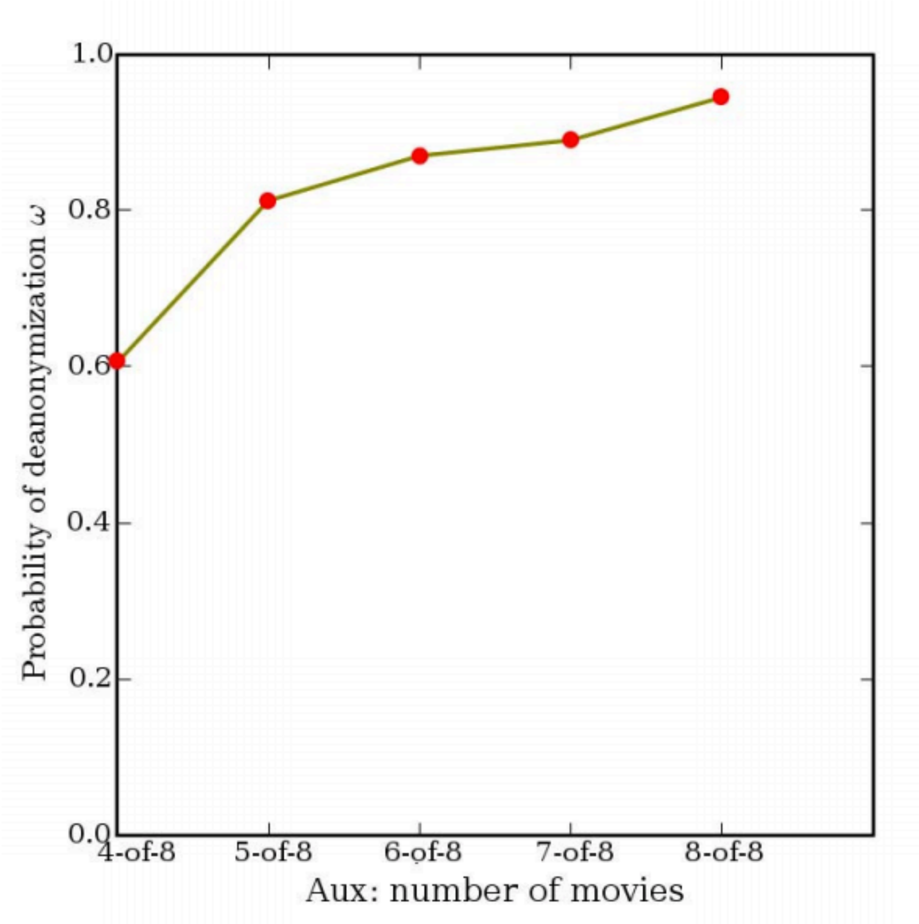


Adversary knows exact ratings but does not know dates at all.



Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (±1) and dates (14- day error).

Empirical results



Effect of increasing error in Aux. in terms of how many movies are correct at all

Roadmap

- ~~Motivation~~
- ~~Privacy definitions~~
- ~~Netflix-IMDb attack~~
- ~~Empirical results~~
- Conclusion

Conclusion

- Naïve anonymization mechanisms do not work

- Even perturbed auxiliary information can be used to launch de-anonymization attacks if:
  - *Database* has many **rare dimensions** and
  - *Auxiliary information* has information about these rare dimensions

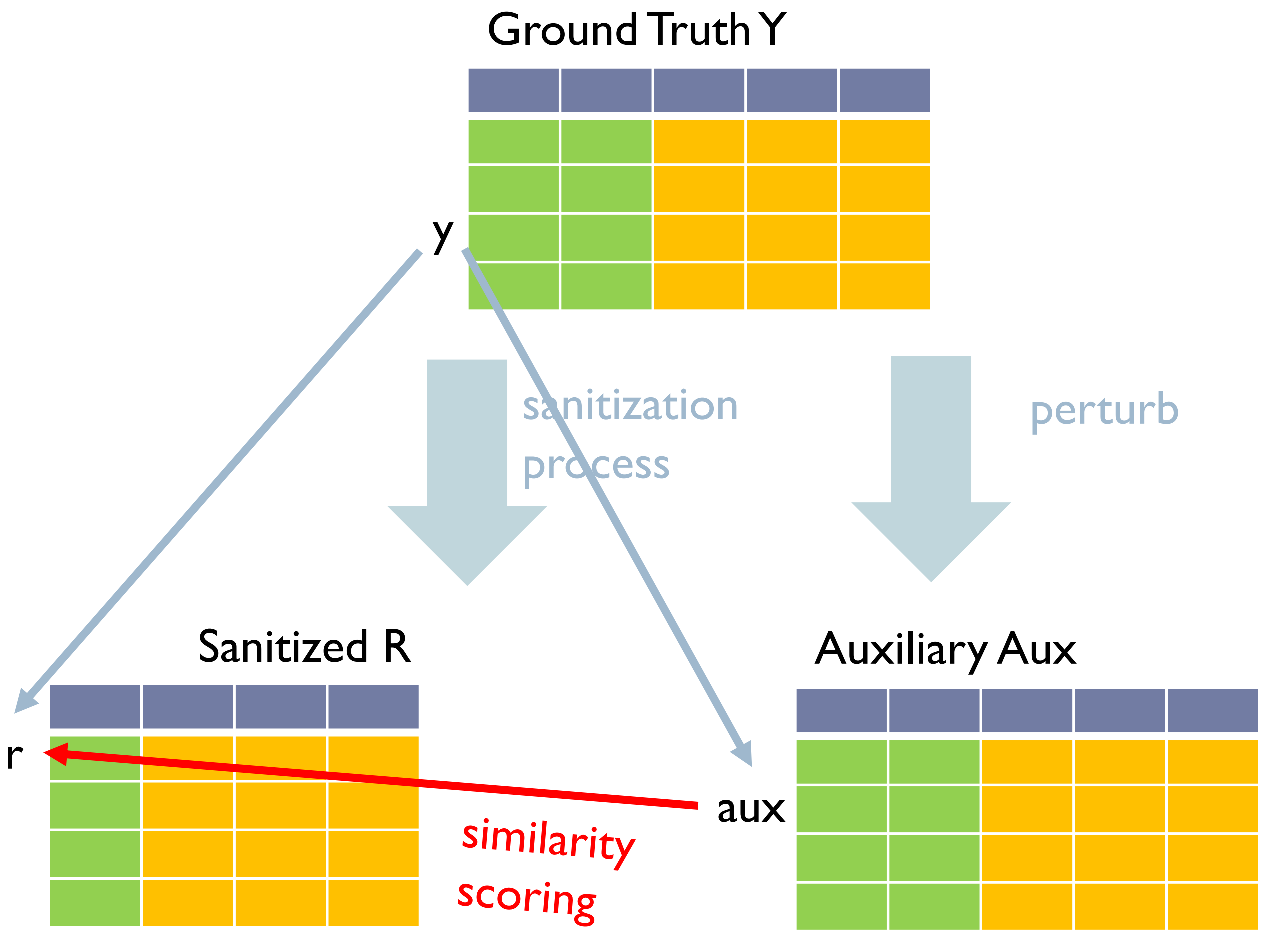
Summary

- Anonymity via sanitization
  - Offline sanitization
  - Online sanitization (next lecture)
- Privacy definitions
  - k-anonymity
  - l-diversity
  - t-closeness
  - m-invariance
  - …

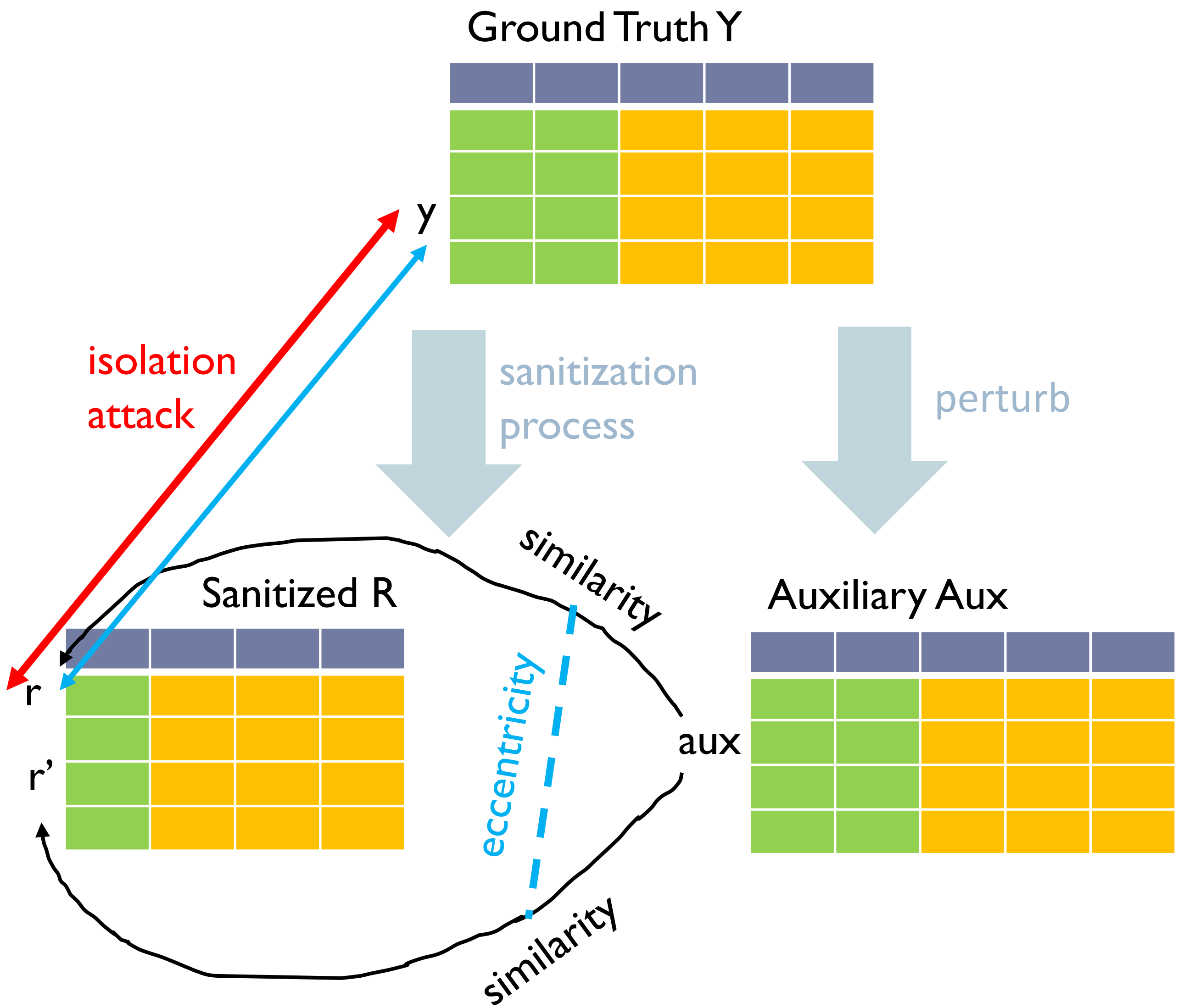
Summary

- Deanonmyization attacks
  - Isolation
  - Amplification
- Measuring attack success without ground truth
  - Measurables
    - similarity
    - eccentricity

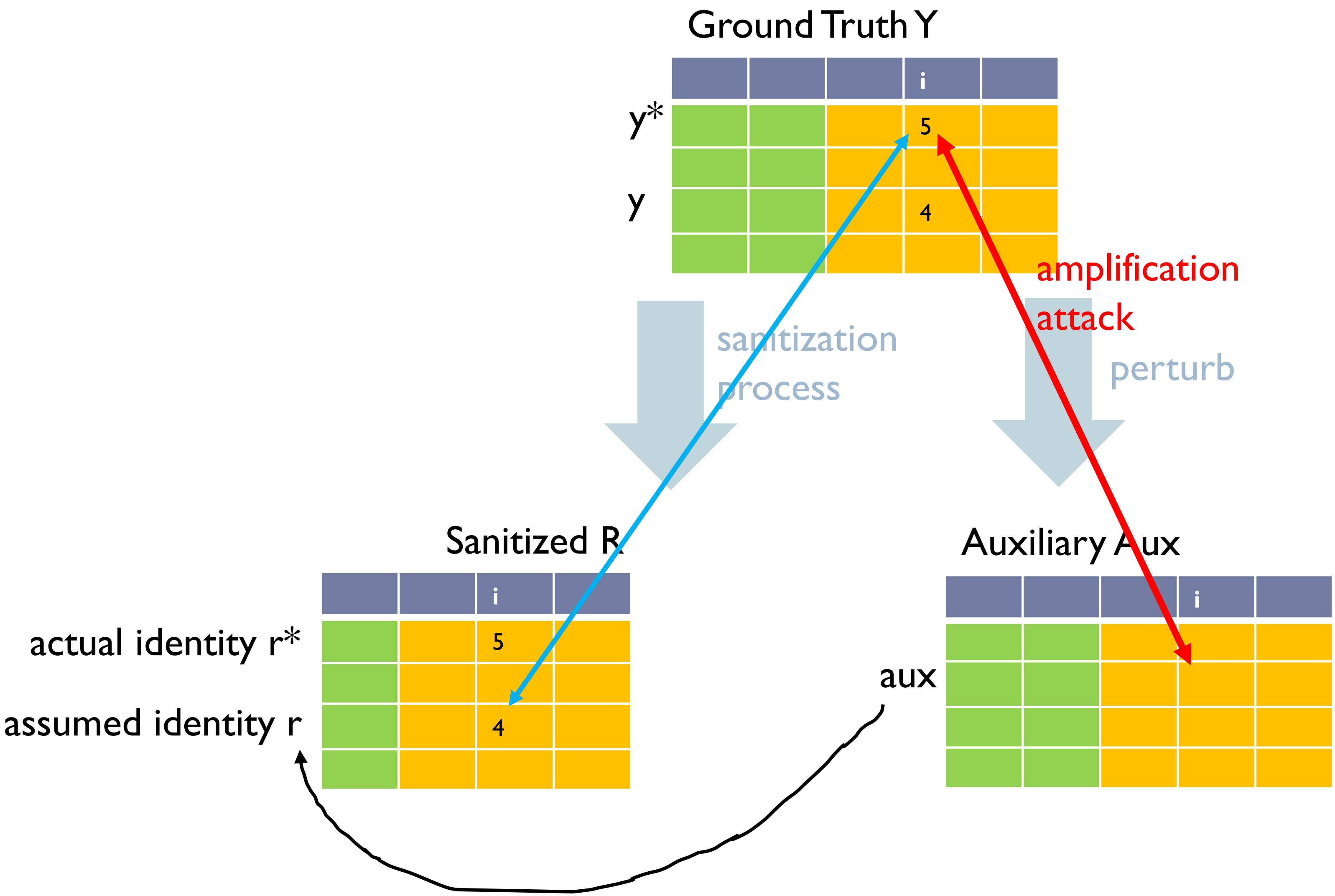
Deanonymization



Isolation attack



Amplification attack



Anonymization settings	
Offline/non-interactive release sanitized dataset	**Online/interactive** sanitize queries
Privacy definitions	
k-anonymity Minimum anonymity set size	**l-diversity** Minimum sensitive range size
T-closeness Minimum variation of distribution of sensitive attribute	

<table>
<tr><th colspan="2">Assumptions and Experimental Measurements
Given aux in Aux, isolate r in D closest to it</th></tr>
<tr><td>Modeling
Y -- Ground Truth records (NOT KNOWN)
R -- Sanitized records
Aux -- Auxiliary records</td><td>Measurements
e – eccentricity
best isolate r vs second best r’</td></tr>
<tr><td colspan="2">Deanonymization attacks</td></tr>
<tr><td>Isolation
Link auxiliary aux in A to r in R.
Is aux is same identity as g.t. y → r ?</td><td>Amplification
Use R to find values of fields not in aux
Are predicted values close to g.t. y ?</td></tr>
</table>