International Workshop on Spatial and Temporal Modeling from Statistical, Machine Learning and Engineering perspectives:STM2016

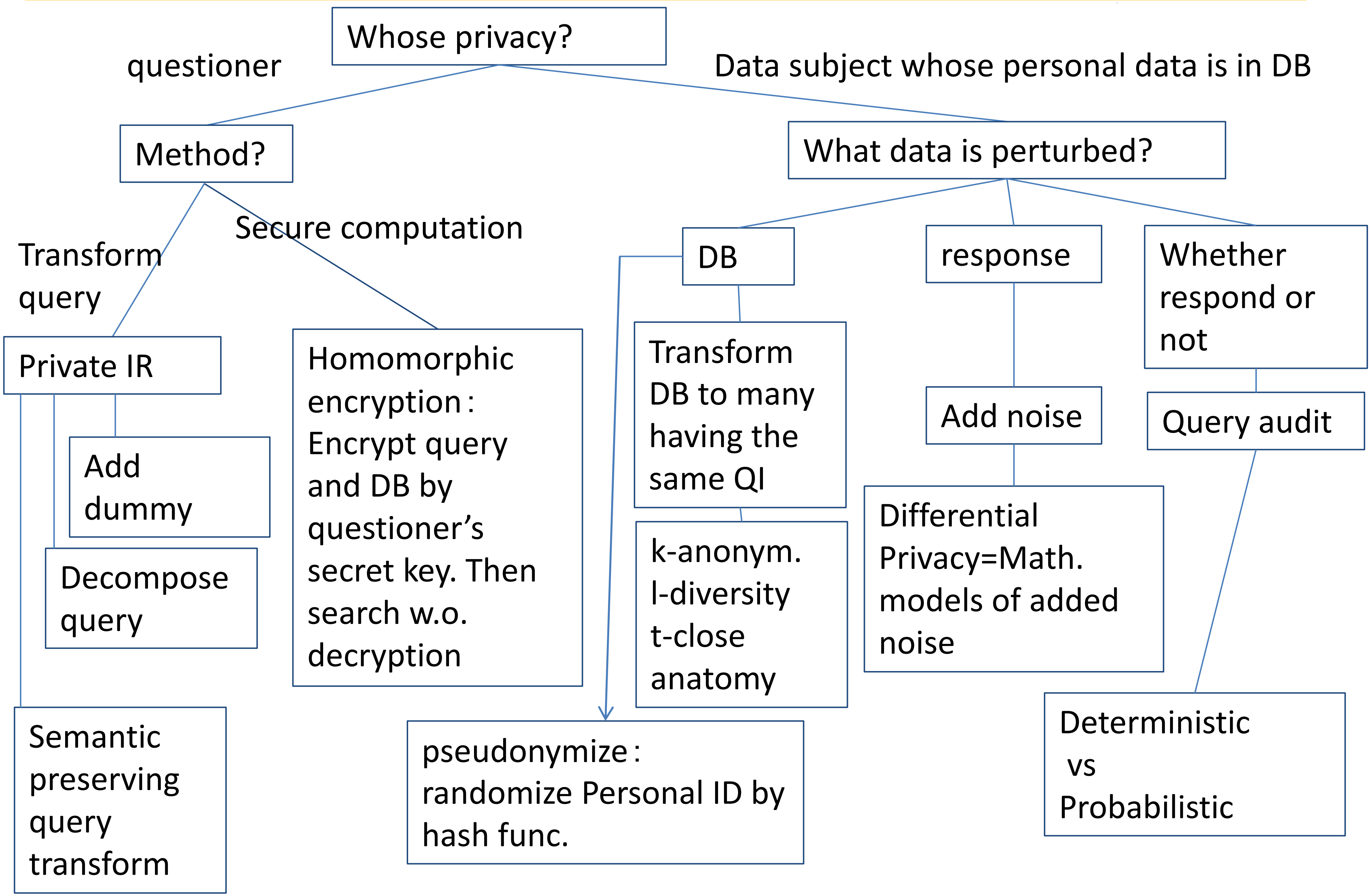
23 July 2016

Privacy Protection：Overview

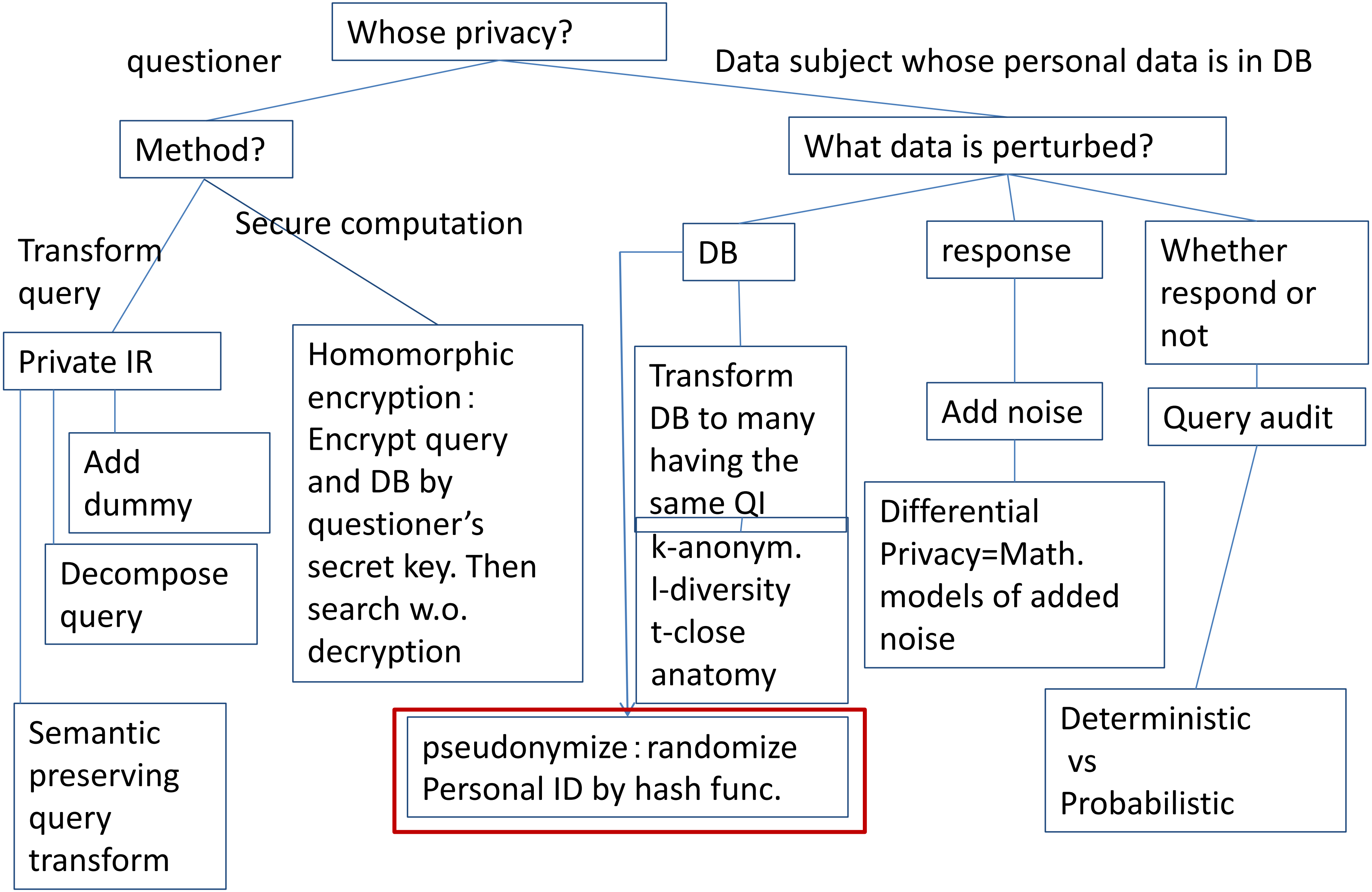
Hiroshi Nakagawa

The University of Tokyo

Overview of Privacy Protection Technologies



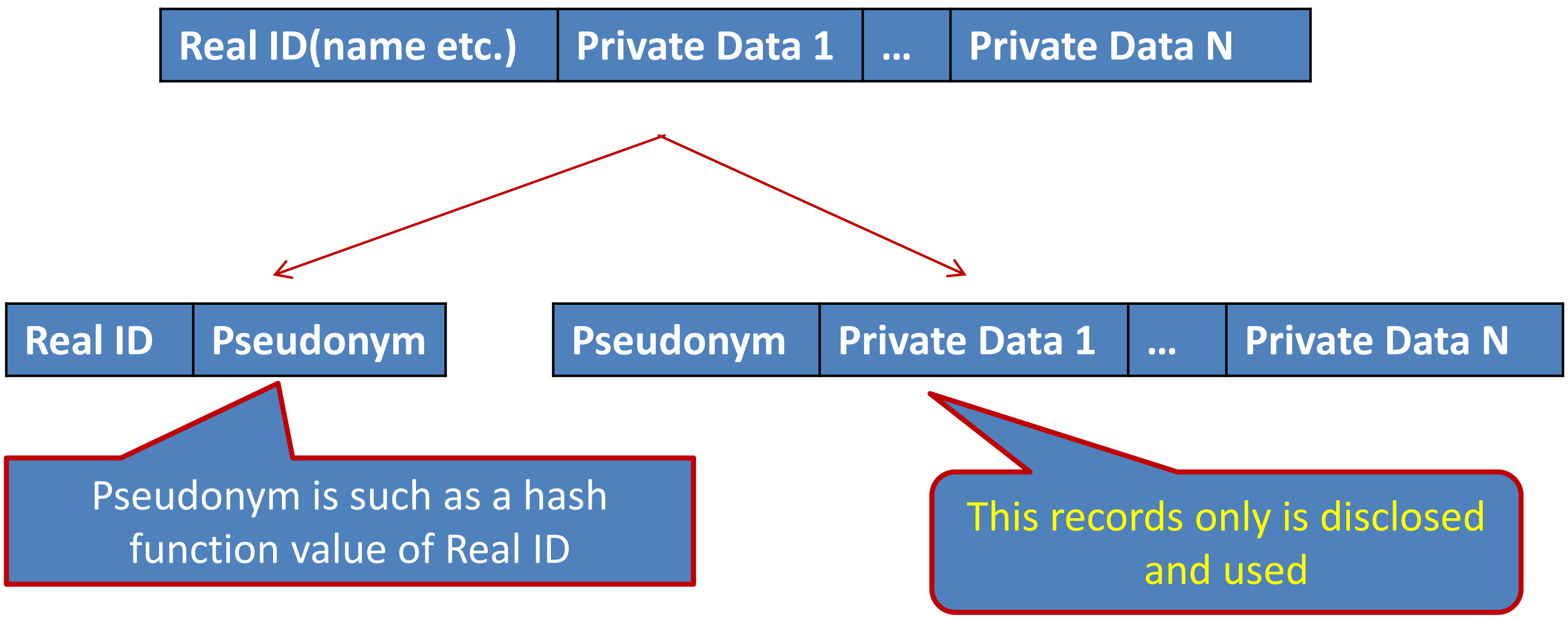
Overview of Privacy Protection Technologies



Updated Personal Information Protection Act in Japan

  - The EU General Data Protection Regulation is finally agreed in 2016
- Japan: Personal Information Protection Act (PIPA): Sep.2015
- Anonymized Personal Information is introduced.
  - Anonymized enough not to de-anonymized easily
  - Freely used without the consent of data subject.
  - Currently, Pseudonymized data is not regarded as Anonymized Personal Information
- Boarder line between pseudonymized and anonymized is a critical issue.

What is pseudonymization?



Variations of Pseudonymization in terms of frequency of pseudonym update

The same individual's personal data

pseu	weight
A123	60.0
A123	65.5
A123	70.8
A123	68.5
A123	69.0

Update pseud.

pseu	weight
A123	60.0
A123	65.5
B432	70.8
B432	68.5
C789	69.0

Frequent update

pseu	weight
A123	60.0
B234	65.5
C567	70.8
X321	68.5
Y654	69.0

Same Info.

weight
60.0
65.5
70.8
68.5
69.0

- No pseudonym update
- Highly identifiable
- Needed in med., farm.

- pseudonym update
- Divide k subsets with different pseudonyms
- Freq. update lowers both identifiability and data value

- Update pseudonym data by data
- Regarded as distinct person's data. No identifiability

obscurity

Is pseudonymization with updating not Anonymized Personal Information (of new Japanese PIPA)?

- Pseudonymization without updating for accumulated time sequence personal data
  - Accumulation makes a data subject be easily identified by this sequence of data
  - Then reasonable to prohibit it to transfer the third party
  - PIPA sentence reads pseudonymized personal data without updating is not Anonymized Personal Information.

- Obscurity, in which every data of the same person has distinct pseudonyms, certainly is Anonymized Personal Information because there are no clue to aggregate the same person's data.

Record Length

pseu	Loc. 1	Loc.2	Loc.3	…
A123	Minato	Sibuya	Asabu	…
A144	Odaiba	Toyosu	Sinbasi	…
A135	…	…	….	….
A526	xy	yz	zw	…
A427				

- No pseudonym update
- High identifiability by long location sequence

transform obscurity

Loc. 1	Loc.2	Loc.3	…
Minato	Sibuya	Asabu	…
Odaiba	Toyosu	Sinbas i	…
…	…	….	….
xy	yz	zw	…

- Even if pseudonym is deleted, long location sequence makes it easy to identify the specific data subject.

Technically, shuffling destroys link between same person's data

Loc. 1	Loc.2	Loc.3	…
Minato	Sibuya	Asabu	…
Odaiba	Toyosu	Sinbasi	…
…	…	….	….
xy	yz	zw	…

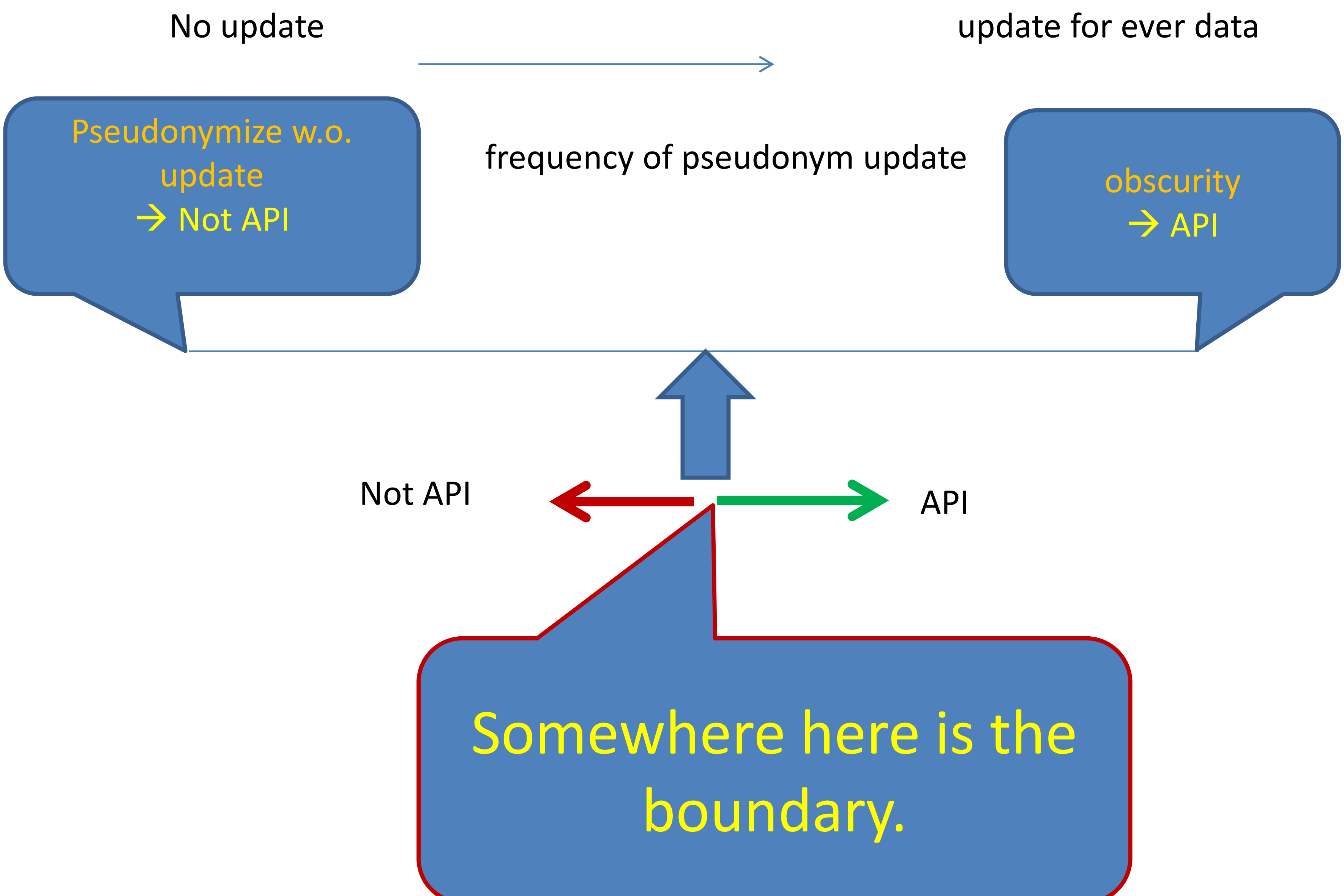
shuffle

Loc. 1	Loc.2	Loc.3	…
Minato	yz	zw	…
Odaiba	Toyosu	Asabu	…
…	…	….	….
xy	Sibuya	Sinbasi	…

obscurity

Almost no clue to identify same individual's record.
But data value is reduced.

The boundary between Anonymized Personal Info.(API) and no API



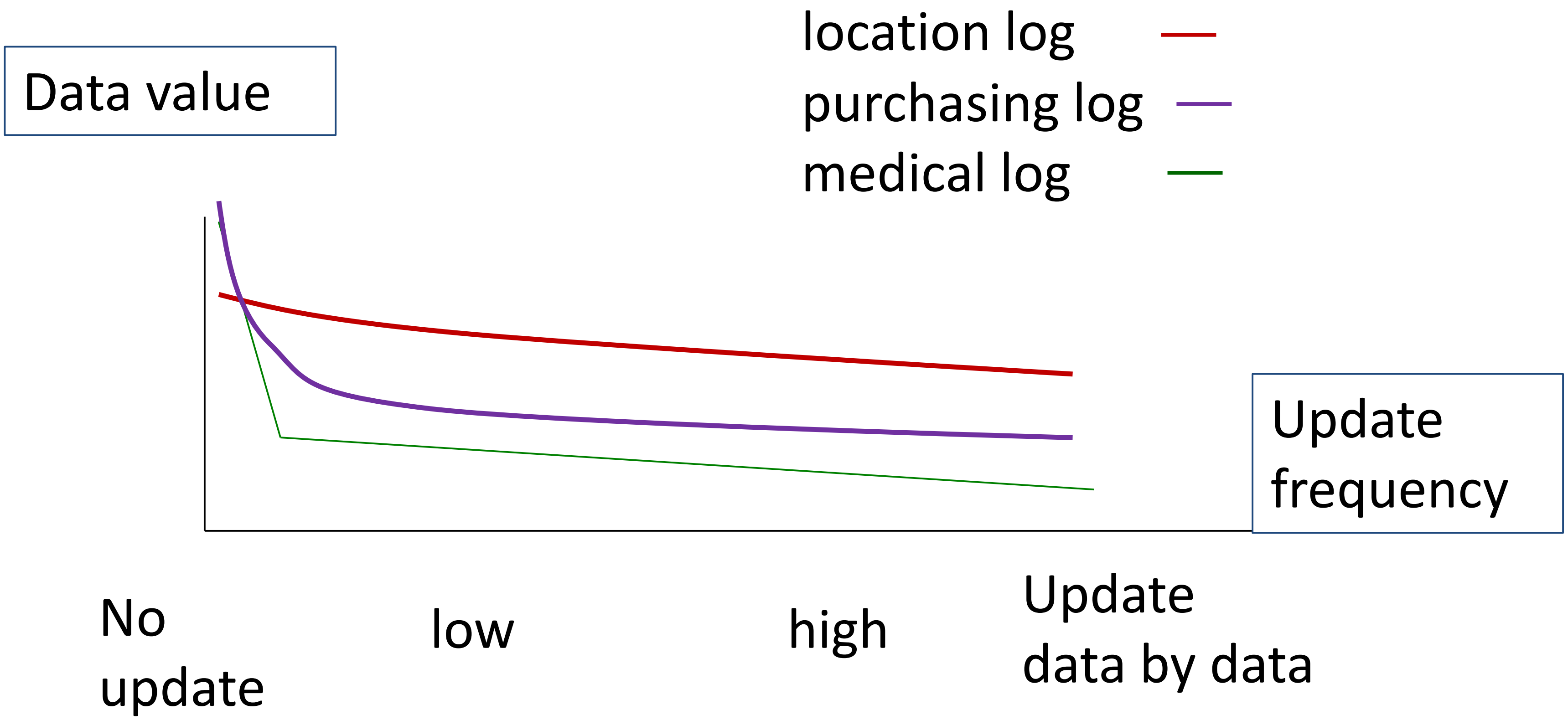
Continuously observed personal data has high value in medicine

- Frequent updating of pseudonym enhances anonymity,
- But reduces data value
  - Especially in medicine.

  - Physicians do not require “no update of pseudonym.”
  - For instance, it seems to be enough to keep the same pseudonym for one illness as I heard from a researcher in medicine.

Updating frequency vs Data value

- see the figure below:



category	Frequency of pseudonym updating	Usage
Medical	No update	Able to analyze an individual patient's log ,especially history of chronic disease and lifestyle
	update	Not able to pursue an individual patient's history. Able to recognize short term epidemic
Driving record	No update	If a data subject consents to use it with Personal ID, the automobile manufacture can get the current status of his/her own car, and give some advice such as parts being in need to repair.
		If no consent, nothing can be done.

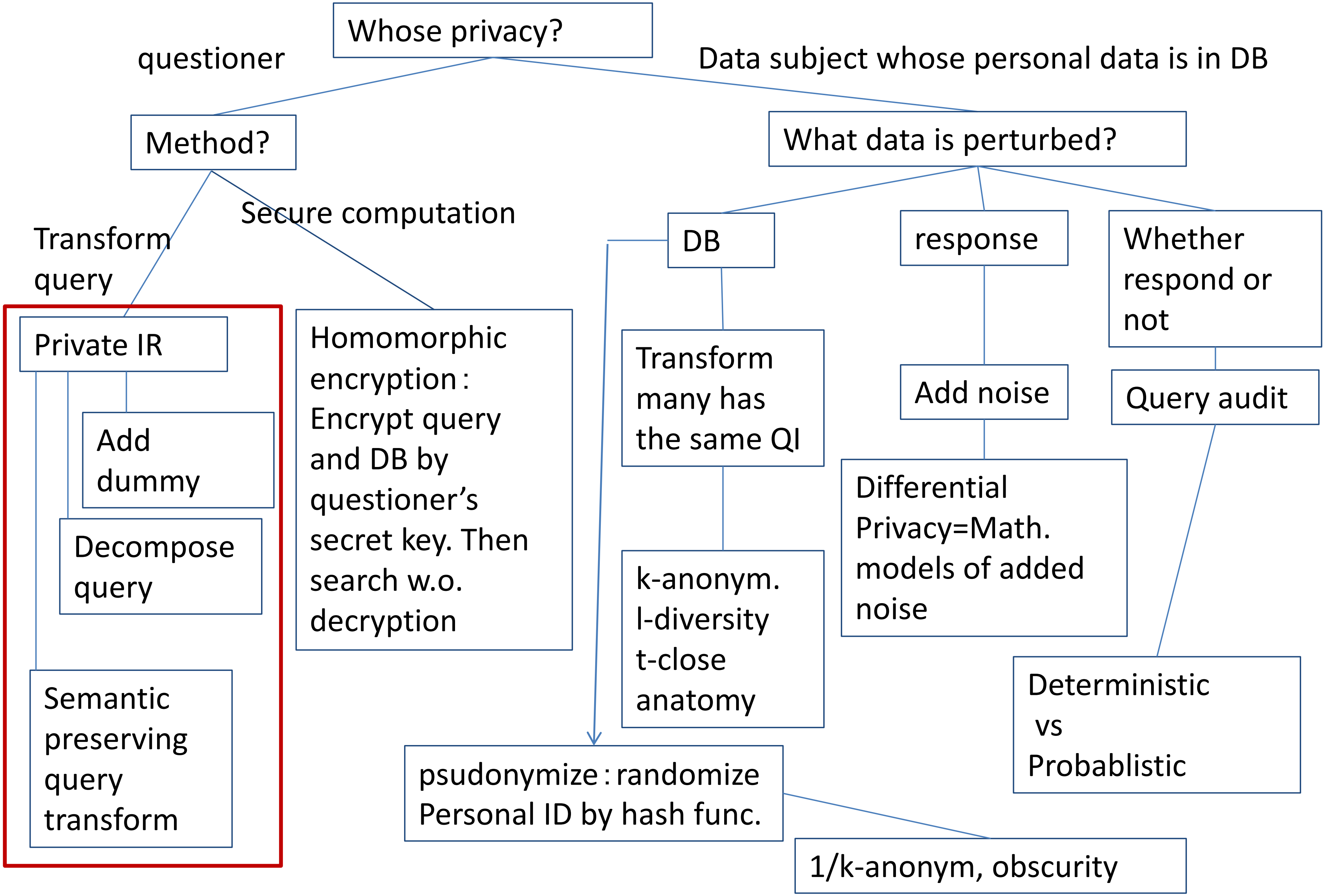
category	Frequency of pseudonym updating	Usage
Driving record	Low frequency	Long range trend of traffic, which can be used to urban design, or road traffic regulation for day, i.e. Sunday.
	High frequency	We can only get a traffic in short period.
Purchasing record	No update	If a data subject consents to use it with Personal ID, then it can be used for targeted advertisement.
		If no consent, we can only use to extract sales statistics of ordinary goods.
	Low frequency	We can mine the long range trend of individual's purchasing behavior.
	High frequency	We can mine the short range trend of individual's purchasing behavior.
	Every data	We only investigate sales statistics of specific goods

Summary: What usage is possible by pseudonymization with/without updating

- As stated so far, almost all psedonymized data are usefull in statistical processing
- No targeted advertisement, nor profiling of individual person

- Pseudonymized data are hard to trace if it is transferred to many organizations such as IT companies.

Overview of Privacy Protection Technologies



Private Information Retrieval (PIR)

what should be kept secret?

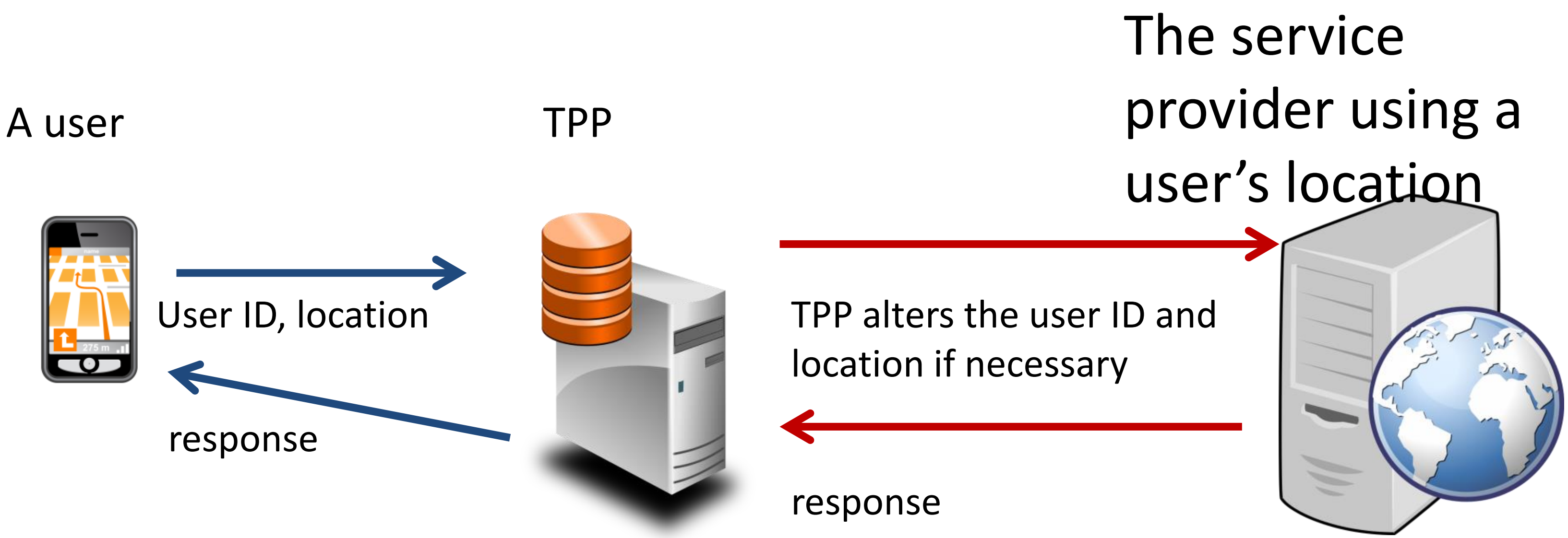
- Information which can identify a searcher of DB or a user of services.
  - Internet ID, name
  - Location from where a searcher send the query
  - Time of sending the query
- Query contents
  - See next slide
- Existence of query

Why user privacy should be protected in IR?

- IT companies in US transfer or even sell user profile to the government authorities such as:
  - AOL responds more than 1000 a month,
  - Facebook responds 10 to 20 request a day
  - US Yahoo sells its members' account, e-mail by 30-40 for one account
- These make amount of profit for IT companies , but no return to data subjects.
  - Even worse, bad guy may steel them.
- Then, internet search engine users should employ technologies that protect him/herself identity from search engine.

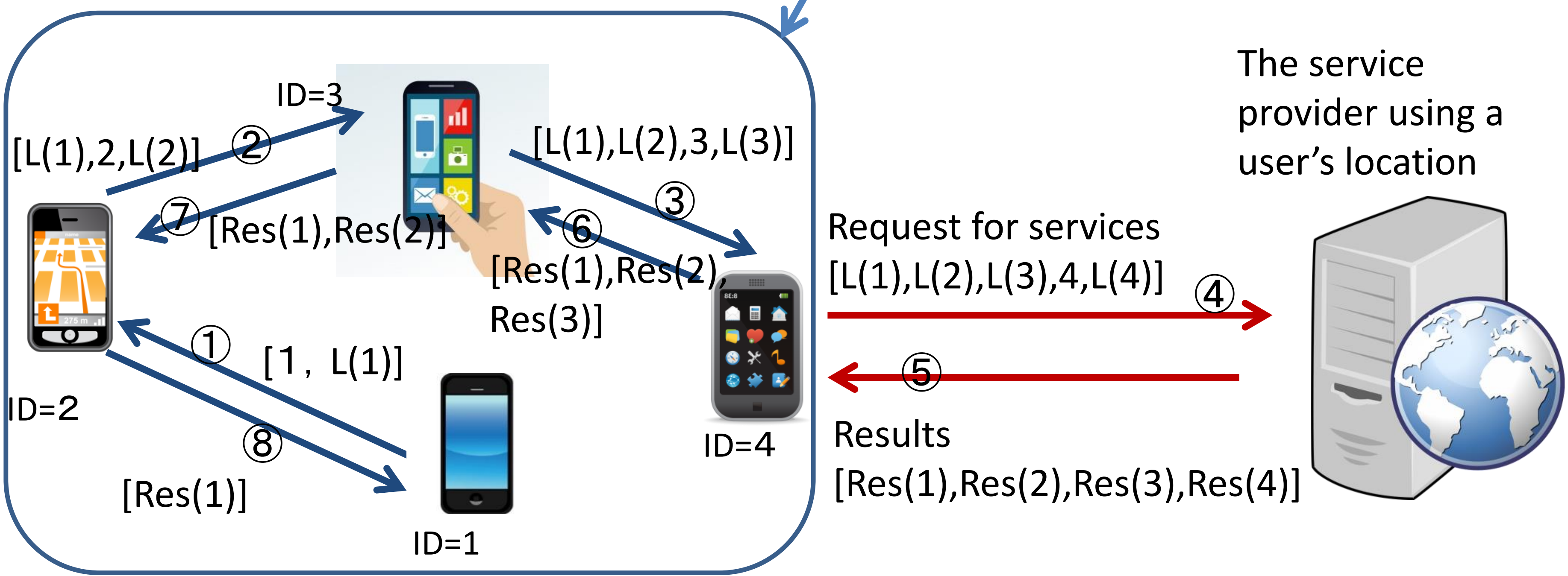
Keep secret the location a user sends a query

- A user wants to use a location based services such as searching near by good restaurants, but does not want the service provider his/her location
- Using the trusted third party :TPP if exists

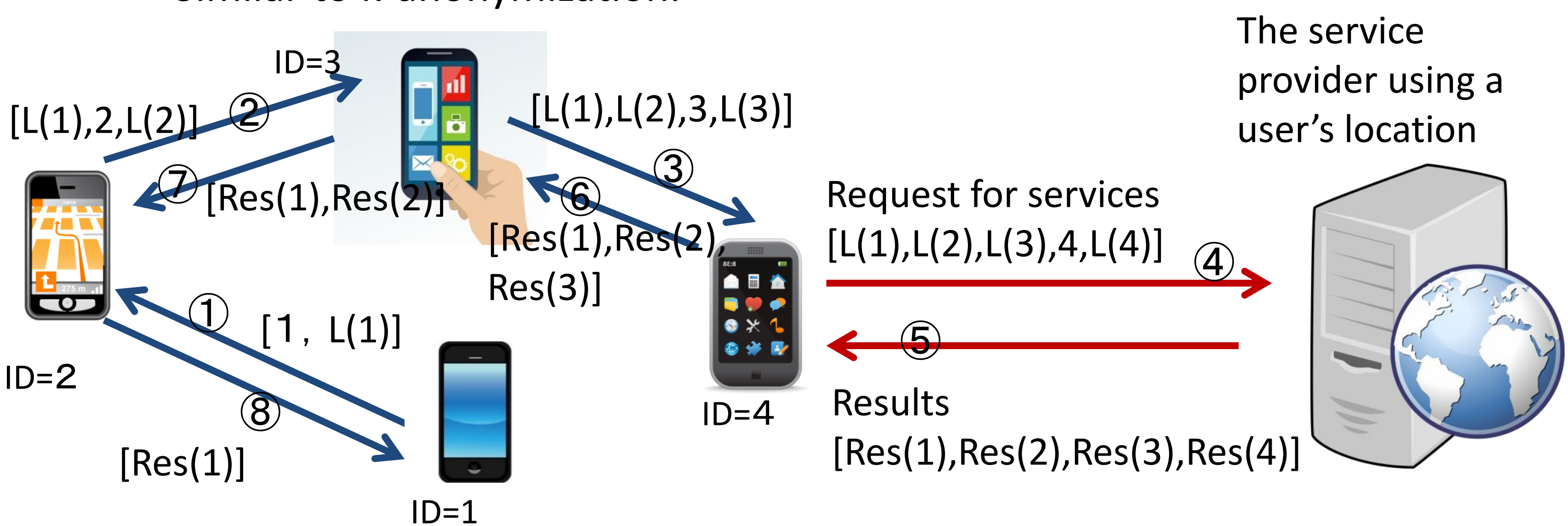


Mixing up several users' locations

- **In case of no TPP, several users trusting each other make <u>a group</u>, and use the location based services**



- L(n) is a location of a user whose ID=n
- Starting from ID=1, and add up each user's location and finally k th user sends the mixed up locations and request the services ①→ ④
- Each user only memorizes the previous user's ID and when receives the response , return it to the previous user as shown in the figure below. ⑤→⑧
  - By shuffling locations in a location list, each user does not recognize which response is for whose request.
  - Similar to k-anonymization.

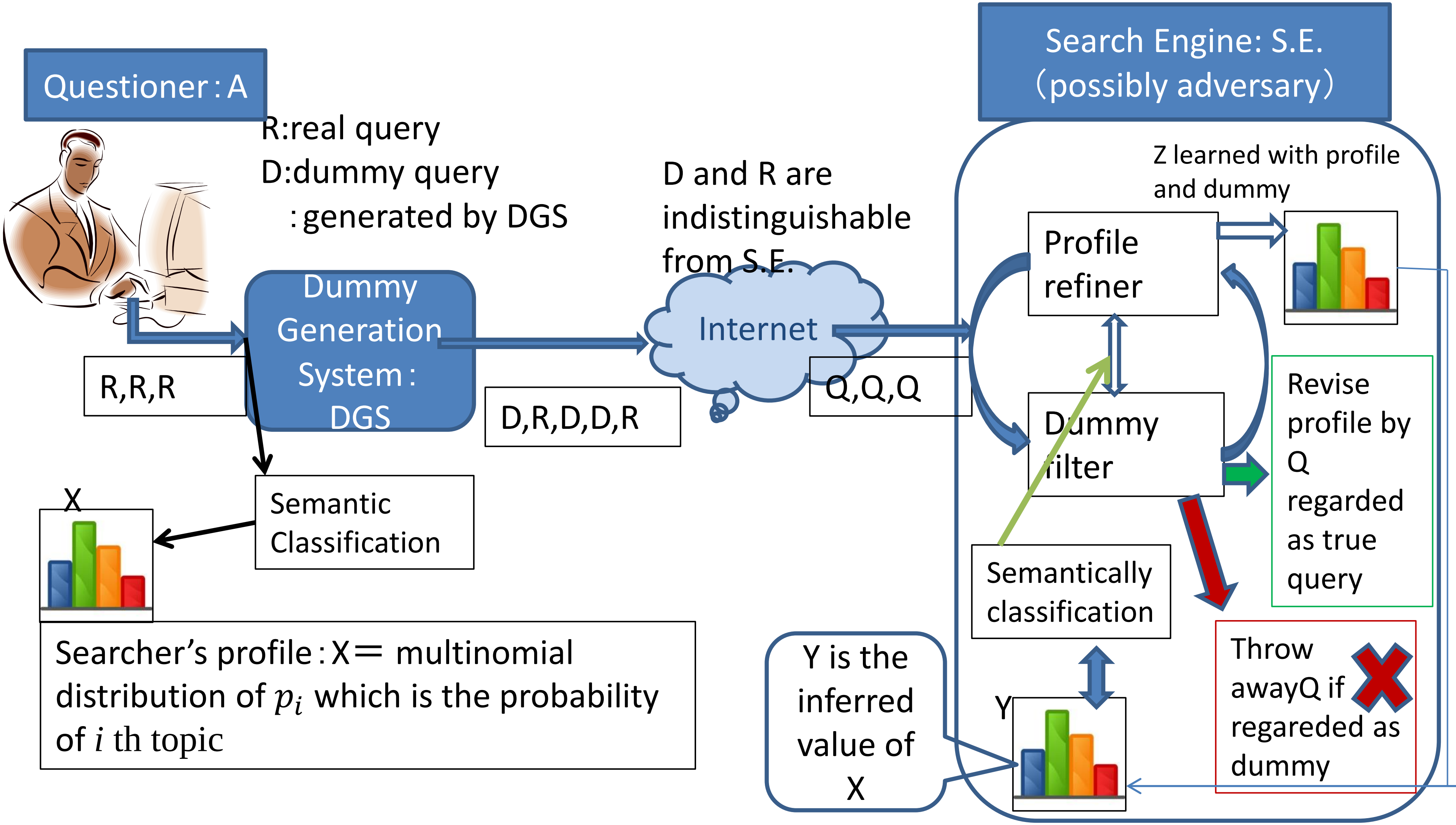


How to make it difficult to infer the real query ? → Obfuscation

- A query is divided into words. Each word is used as distinct query
- Add dummy term, say confusing words, to the query
- Replace a query word with semantically similar word(s)

➢ When we get response(list of documents, etc.), we have to select out the originally intended answer from them.

Outlook of PIR with obfuscation



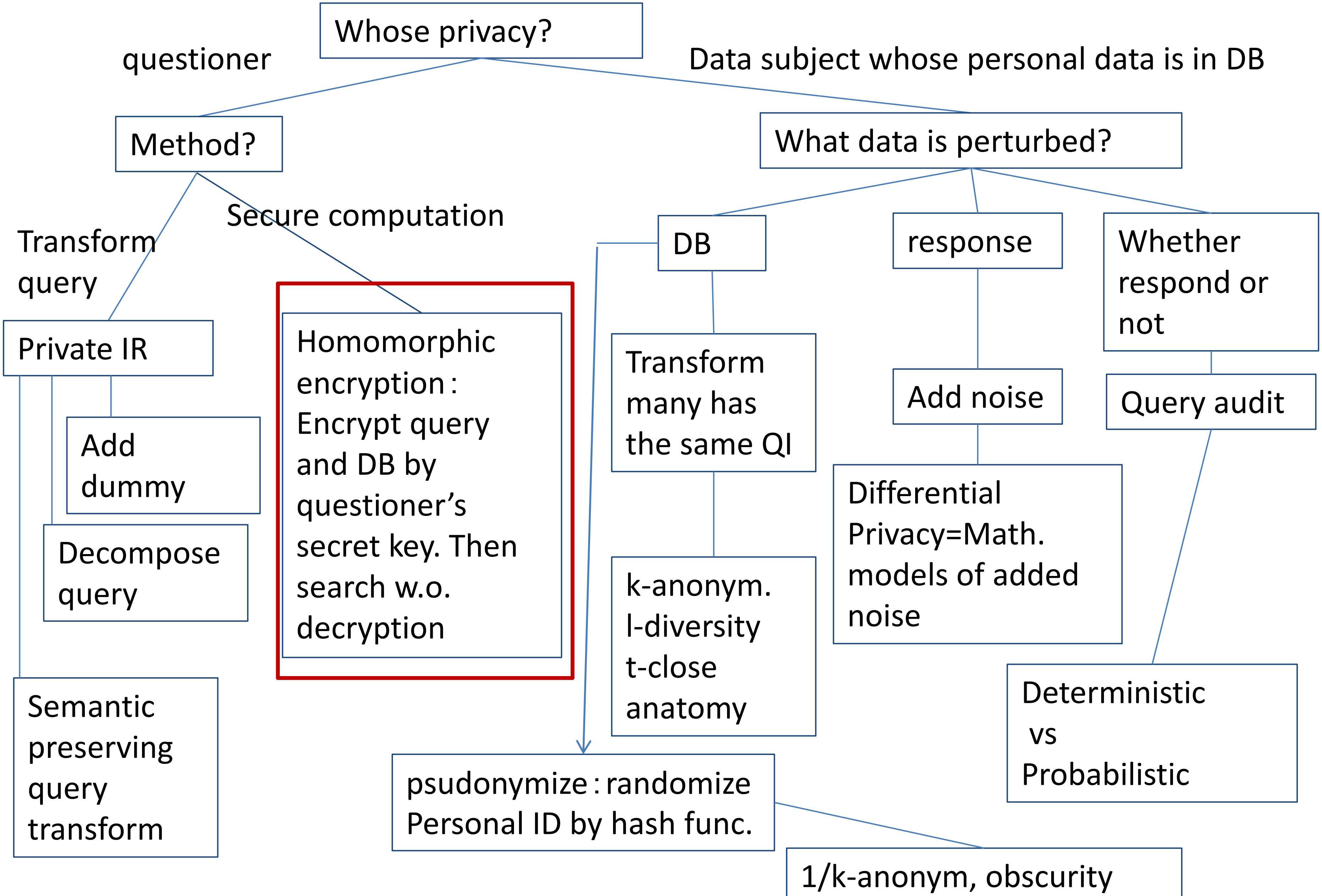
Supplemental explanation

- A questioner : A makes dummy queries D by DGS(dummy generater system) based on the real query R, and send R and D to the search engine: S.E., which might be an adversary.

- S.E. receives Q which actually consists of R and D. Then S.E. learns a questiner's profile Z, and classifies Q into real query and dummy queries.

- In this setting, the questioner wants Q not be classified into R and D. In addition, he/she would not like his/her profile inferred by S.E.. That is why adding D or replacing true R with other words.

Overview of Privacy Protection Technologies



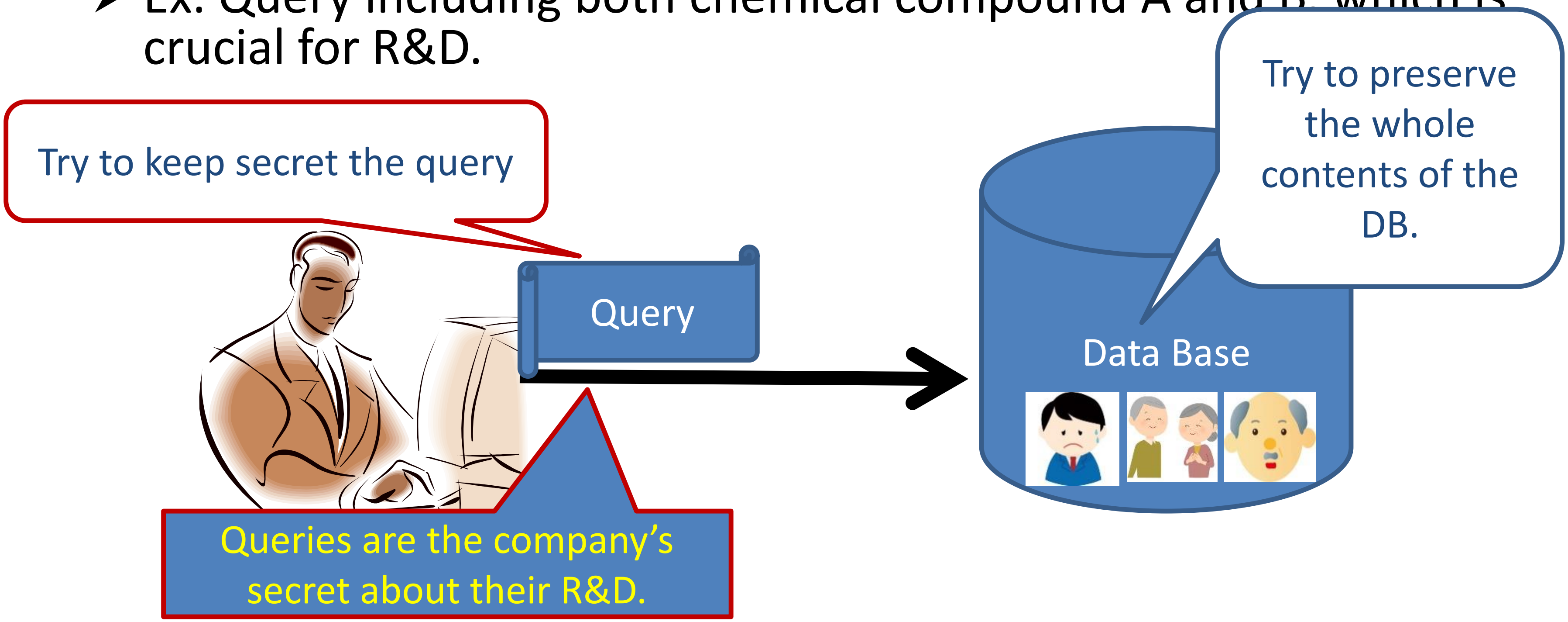
IR with Secure Computation

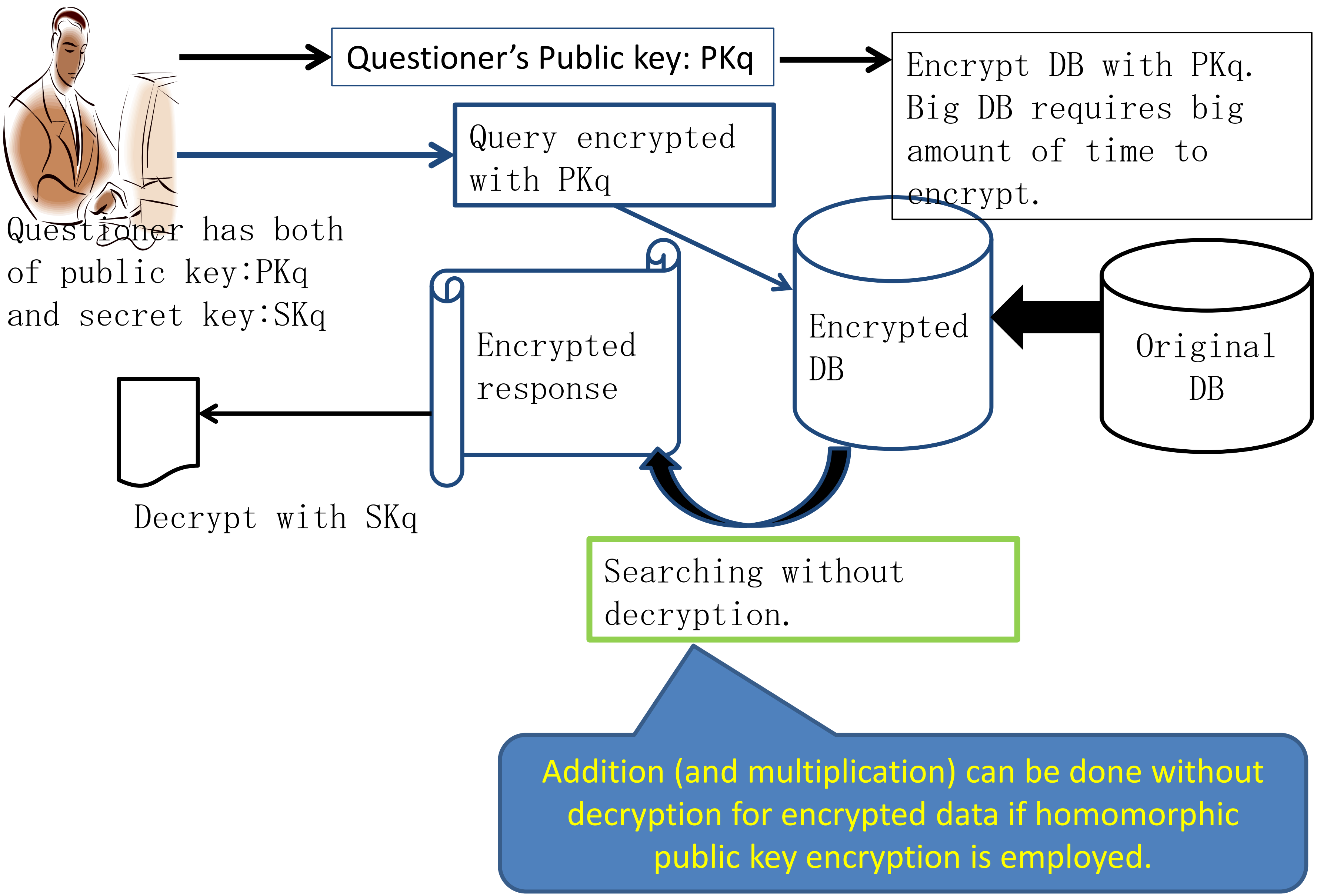
Private Information Retrieval

- Researchers in industry send queries to S.E. to search the DB. Their queries indicate the information of R&D of their company.

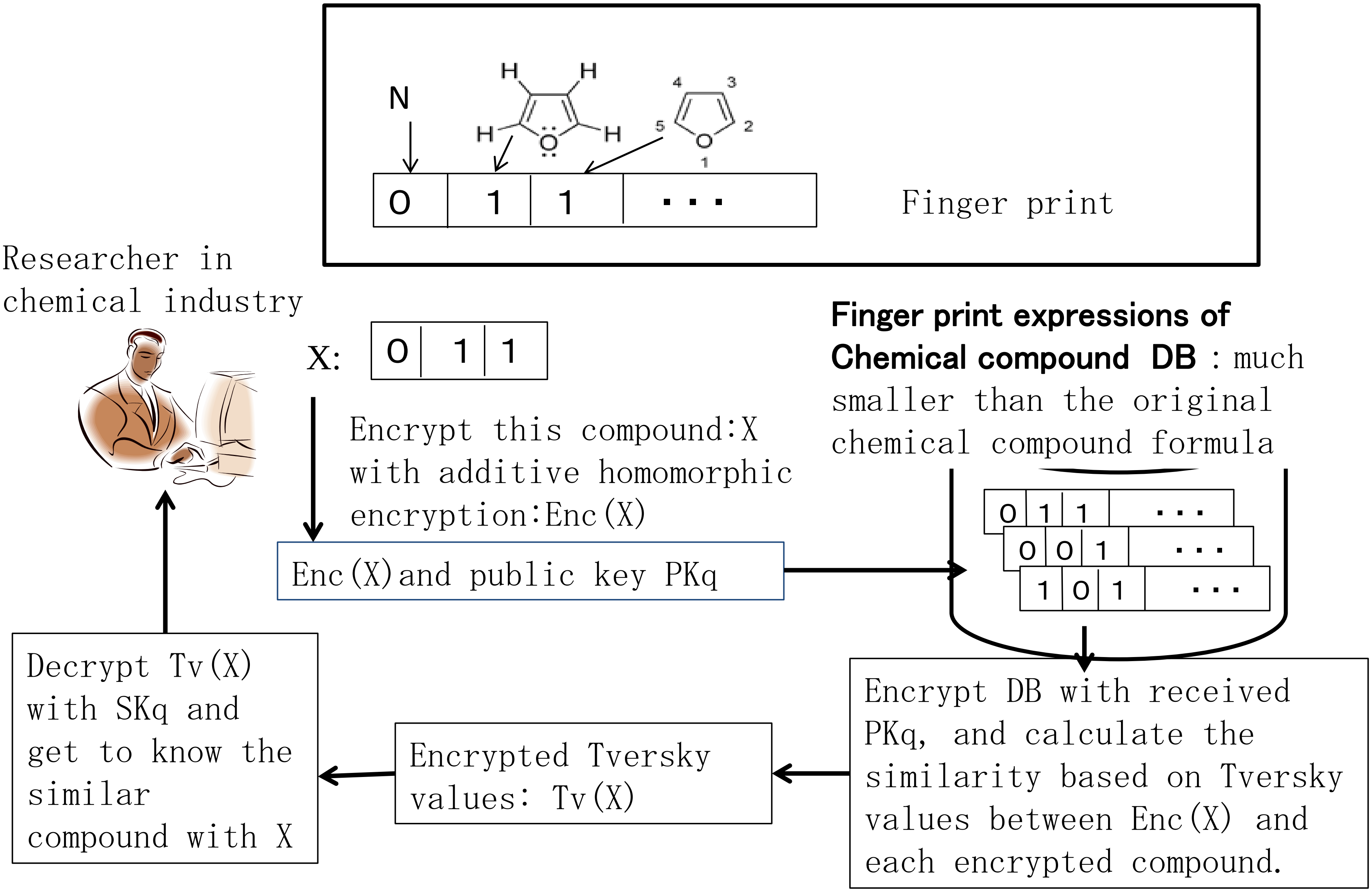
- They want to make the queries secret from S.E. of the DB.

  - Ex. Query including both chemical compound A and B, which is crucial for R&D.

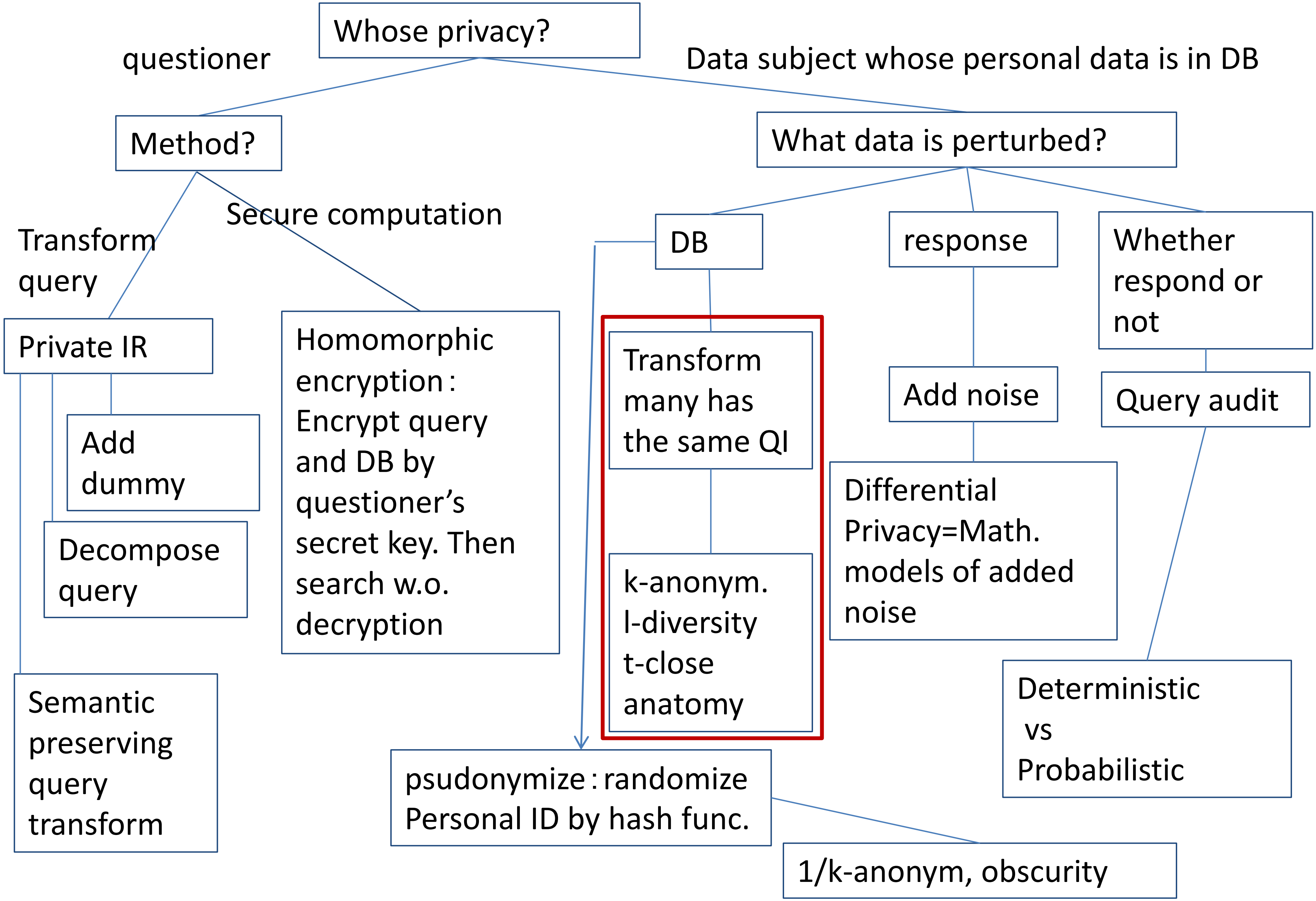



Questioner's Public key: PKq
Encrypt DB with PKq. Big DB requires big amount of time to encrypt.
Query encrypted with PKq
Questioner has both of public key:PKq and secret key:SKq
Encrypted response
Encrypted DB
Original DB
Decrypt with SKq
Searching without decryption.
Addition (and multiplication) can be done without decryption for encrypted data if homomorphic public key encryption is employed.

Chemical Compounds IR based on Secure Computation: Developed by AIST Japan



Overview of Privacy Protection Technologies



k-anomymity, l-diversity

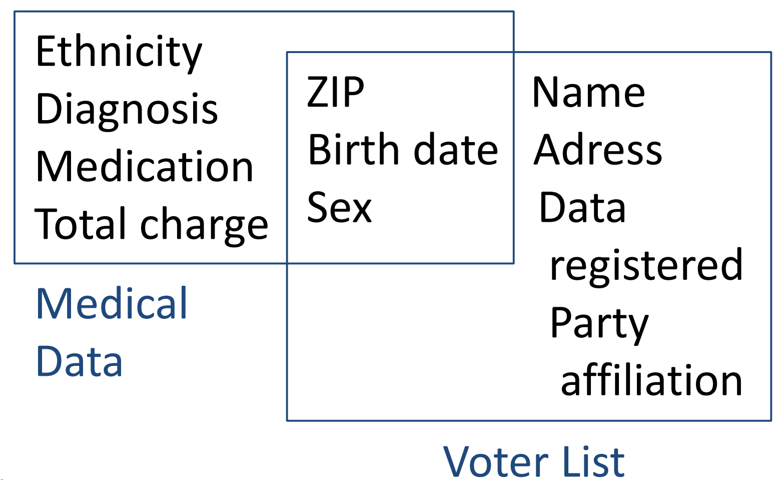
motivation

- Can we anonymize personal data only by removing invididual ID such as name and exact address?
- No
  - Private information can be inferred by combining the publicly open data: Link Attack
    - Un-connetable anonymity in Japanese medicine mainly for research purpose: Pseudonymize and delete the linking data between psedonym and personal ID.
      - If the linking data is not deleted, we call “Connetable anonymity.”
    - Un-connetable anonymity is thought to be protecting patients’ personal medical data because this kind of data are only confined in the medical organization.
    - If, however, the patients’ data are used in nursing care organization or medicine related companies such as pharmaceutical companies.

Classic Example of Link

- Sweeney [S01a] said the governor of Massachusetts William Weld ‘s medical record was identified by linking his medical data which deletes his name, and the voter as shown in the figure.

- Combining both database
  - 6 people have the same birth date of the governor
  - Within these 6 people, three are male.
  - Within these three, only one has the same ZIP code!



- According to the US 1990 census data,
  - 87% of people are uniquely identified by zipcode, sex, and birth

➢ K-anonymization was proposed to remedy this situation.

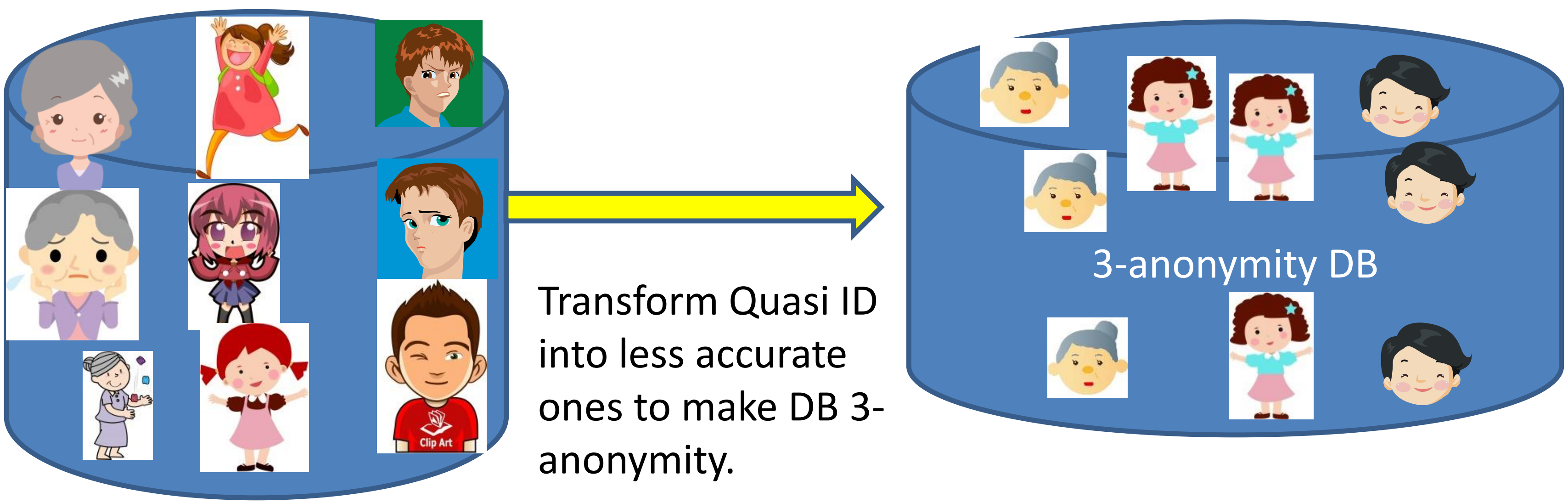
k-anonymity

- Two methods to protect personal data stored in databases from link attacks when this database is transferred or sold to the third party.

  - Method１:　Only Randomly sampled personal data is transferred because whether specific person is stored in this sample DB or not is unknown.

  - Method2:　Transform Quasi ID (address, birthdate, sex) less accurate ones in order that at least k people has the same less accurate Quasi ID: k-anonymization.

  - In the right DB of the figure below, 3 people has the same (less accurate) Quasi ID, say old lady, young girl, young boy →3-anonymity



Example of transforming Quasi ID less accurate

- Attribute of Quasi ID
  - Personal ID（explicit identifiers）is deleted: anonymize
  - Quasi ID can be used to identify individuals
  - Attribute, especially sensitive attribute value should be protected

delete

Personal ID	Quasi ID			Sensitive info.
~~name~~	Birth date	gender	Zipcode	Disease name
~~John~~	21/1/79	M	53715	flu
~~Alice~~	10/1/81	F	55410	pneumonia
~~Beatrice~~	1/10/44	F	90210	bronchitis
~~Jack~~	21/2/84	M	02174	sprain
~~Joan~~	19/4/72	F	02237	AIDS

The objective : Keep each individual identified by Quasi ID

Example of k-anonymity

Original DB

Birth day	gender	Zipcode
21/1/79	M	53715
10/1/79	F	55410
1/10/44	F	90210
21/2/83	M	02274
19/4/82	M	02237

2-anonymized DB

	Birth day	gender	Zipcode
group 1	*/1/79	human	5****
	*/1/79	human	5****
suppress	~~1/10/44~~	~~F~~	~~90210~~
group 2	*/*/8*	M	022**
	//8*	M	022**

Terminology: identify, specify

- Just the summary of basic terminology in Japanese

- specify : A data record becomes known to match to the real world uniquely specified natural person by linking an anonymized personal DB and other non anonymized personal DB

- Identify (or single out) : Data records of several DBs, are known to be the unique same person's data record by linking Quasi ID of these DBs

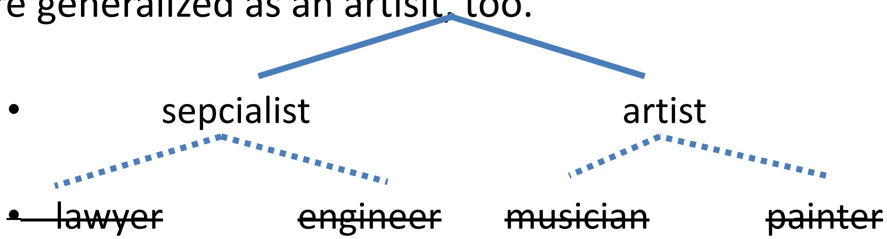
- Without identified, specification is generally hard
  - Neither identified nor specified case: Non-identify&non-specify
  - Identified but not specified: Identify&non-specify

k-anonymization

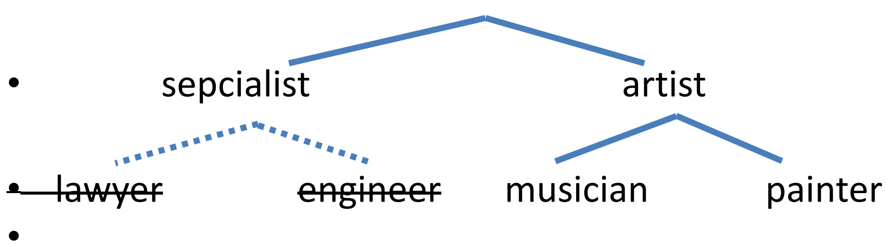
- Sweeney and Samarati [S01, S02a, S02b]
- k-anonymization: transform quasi IDs to less accurate ones so that at least k people have the same quasi IDs.
  - By k-anonymization, the probability of being identified becomes less than **1/k** against link attack.
- Method
  - Generalization of quasi ID values, or suppress a record having a certain value of quasi ID.
    - Not adding noise to attribute value
- Notice the tradeoff between privacy protection and data value degradation (especially for data mining)!
  - Don’t transform more than necessary for k-anonymity!

Generalizations (1)

- Every node of the same level of classification tree are generalized as shown in the figure below:
- Global generalization → accuracy downgraded a lot
  - If a lawyer and an engineer are generalized as a specialist, then a musician and a painter are generalized as an artisit, too.



- Only generalizing nodes in the subtree
  - Even if a lawyer and an engineer are generalized as a specialist, a musician and a painter are not generalized. Avoiding non-necessary generalization.



Generalizations (2)

- Only one of children in a subtree is generalized

  - specialist artist

  - ~~lawyer~~ engineer musician painter

- Local generalization :
  - not all records but individual records are generalized .
  - Good point is less accuracy reduction.
    - i.e. John(lawyer) → John(specialist) but Alex(lawyer) still remains a lawyer.

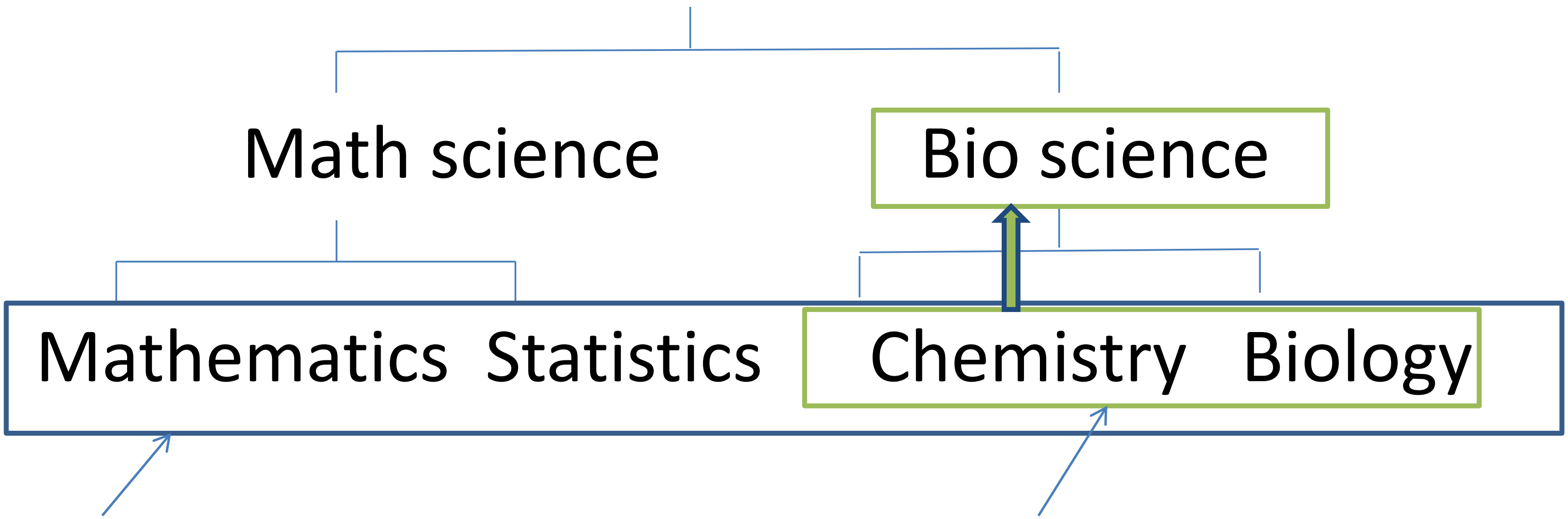
Evaluation function in k-anonymization

- K-anonymization algorithm uses the following evaluation function to control whether generalization continues or stop.

- minimal distortion metric:*MD*
  - The number of lost precise data by generalization.
  - For example, 10 engineers are generalized into specialist, MD=10

- $ILoss(v_g) = \frac{|v_g|-1}{|D_A|}$: The loss when more precise data than v_g is generalized to v_g
-

  $|v_g|$ is the number of kinds of data of v_g's children.
  $|D_A|$ is the number of kinds of data of $v_g' s$ attribute:A



$|D_A|$=4

$|v_g|$=2

$$ILoss(v_g) = \frac{|v_g| - 1}{|D_A|} = \frac{2 - 1}{4} = \frac{1}{4}$$

- Trade-off between information accuracy and privacy

- $IGPL(s) = \frac{IG(s)}{PL(s)+1}$
  - s means generalizing to data
  - $IG(s)$ is the loss of information gain or MD by applying s
  - $PL(s)$ is the degree of anonymization by applying s
    - If k-anonymization, the degree is k.

Lattice for generalization k-anonymity

Z2 ={537**}

↑

Z1 ={5371*, 5370*}

↑

Z0 ={53715, 53710, 53706, 53703}

zipcode

B1 ={*}

↑

B0 ={26/3/1979, 11/3/1980, 16/5/1978}

Birth date

S1 ={Person}

↑

S0 ={Male, Female}

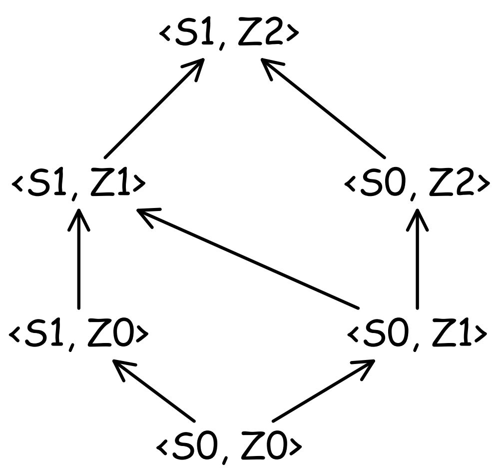
sex

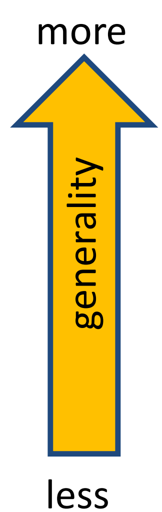
Objective

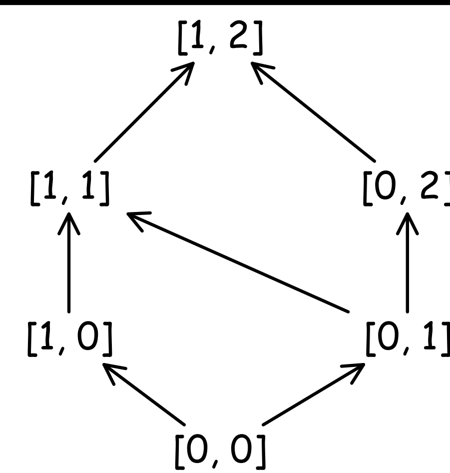
Minimum generalization

Subject to k-anonymity

Lattice for generalization of all quasi IDs



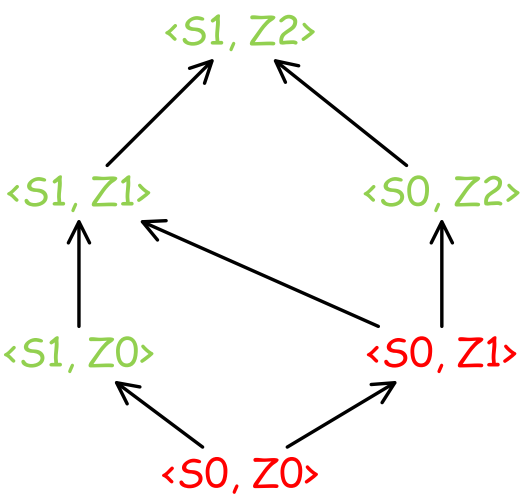




Use lattice for efficient generalization

incognito [LDR05]

Using monotonicity



To simplify, only about <S,Z>

(I) Generalization property (~rollup)

if k-anonymity at a node

then nodes above the node satisfy k-anonymity

e.g., <S1, Z0> satisfies k-anonymity

→ <S1, Z1> and <S1, Z2> satisfy k-anonymity

(II) Subset prpperty (~apriori)

if a set of quasi ID does not satisfy k-anonymity at a node

then a subset of the set of quasi ID does not satisfy k-anonymity

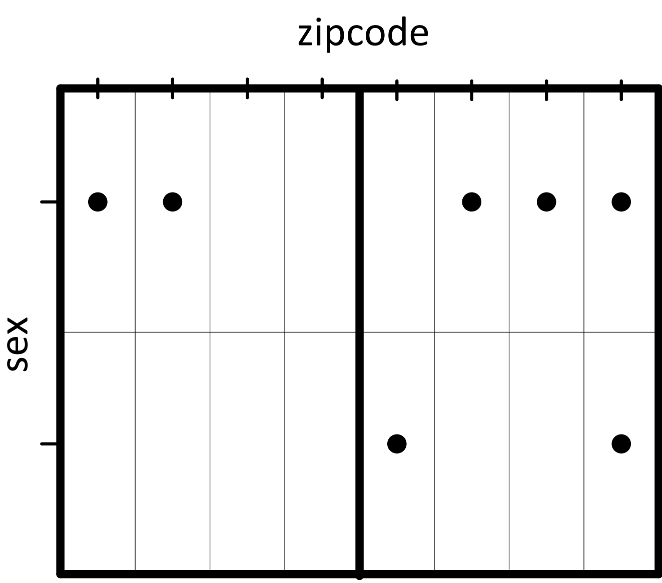
e.g., <S0, Z0> k-匿名性 でない

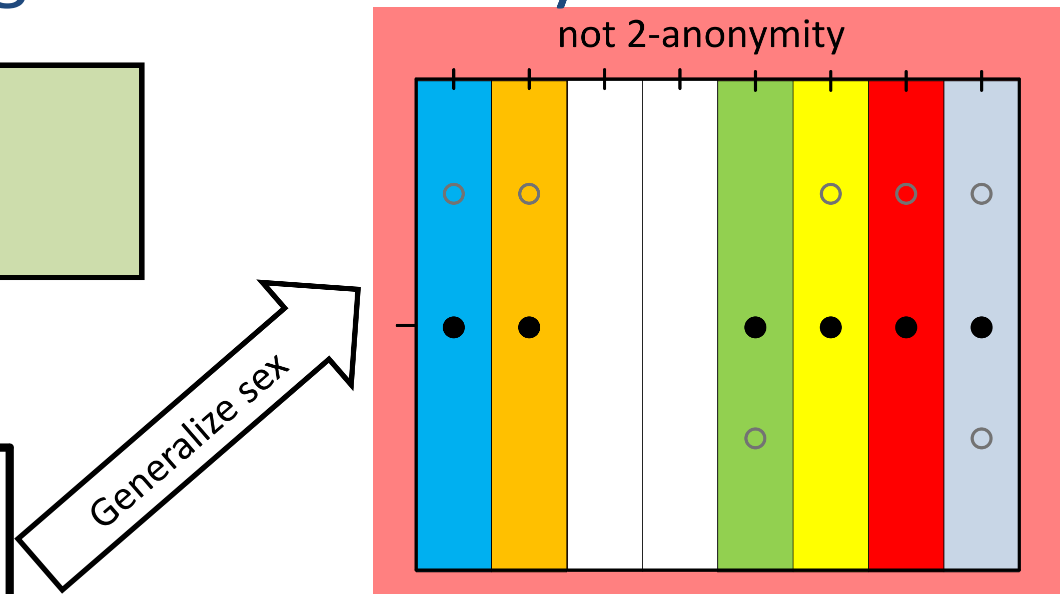
→ <S0, Z0, B0> and <S0, Z0, B1> k-匿名性 でない

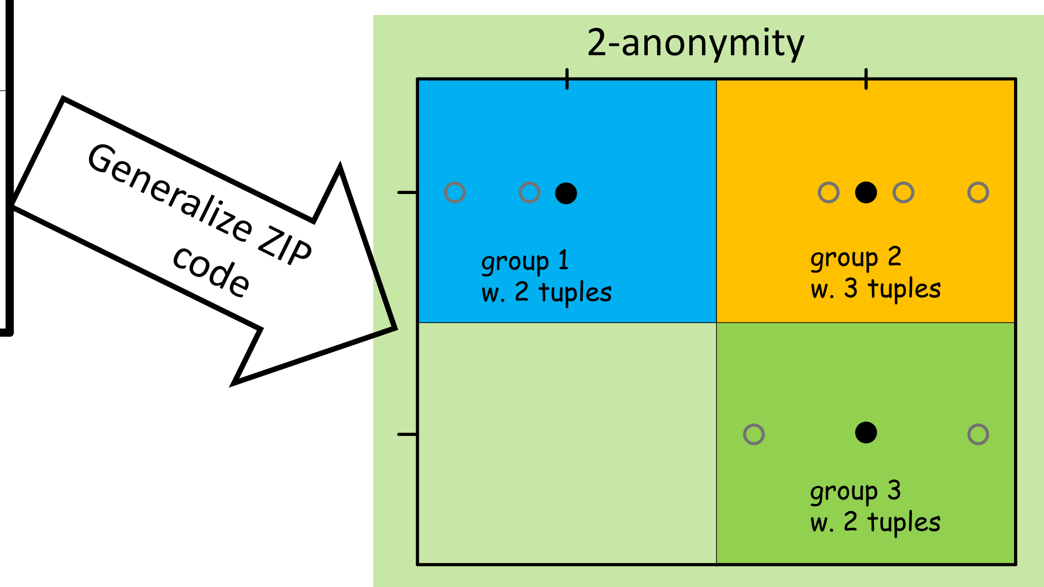
Example Case:
Dividing does not anonymize

Example of Incognito

2 quasi ID , 7 data point

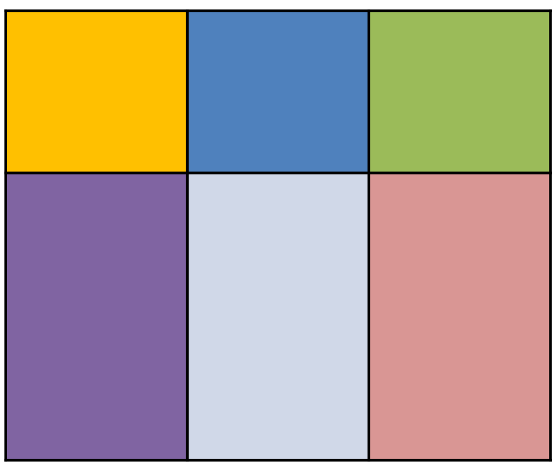




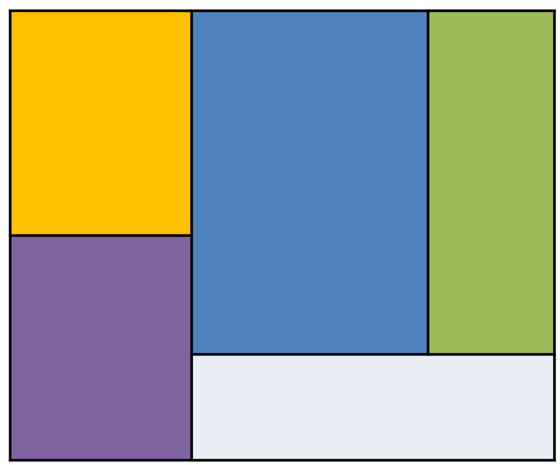


Examples [LDR05, LDR06]

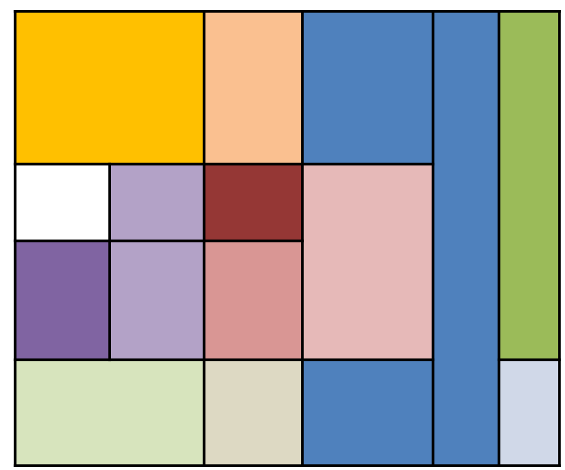
Each dimension is sequentially generalized
incognito [LDR05]

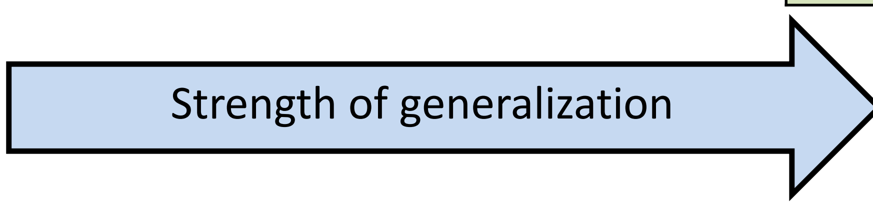


Each dimension is independently generalized
mondrian [LDR06]



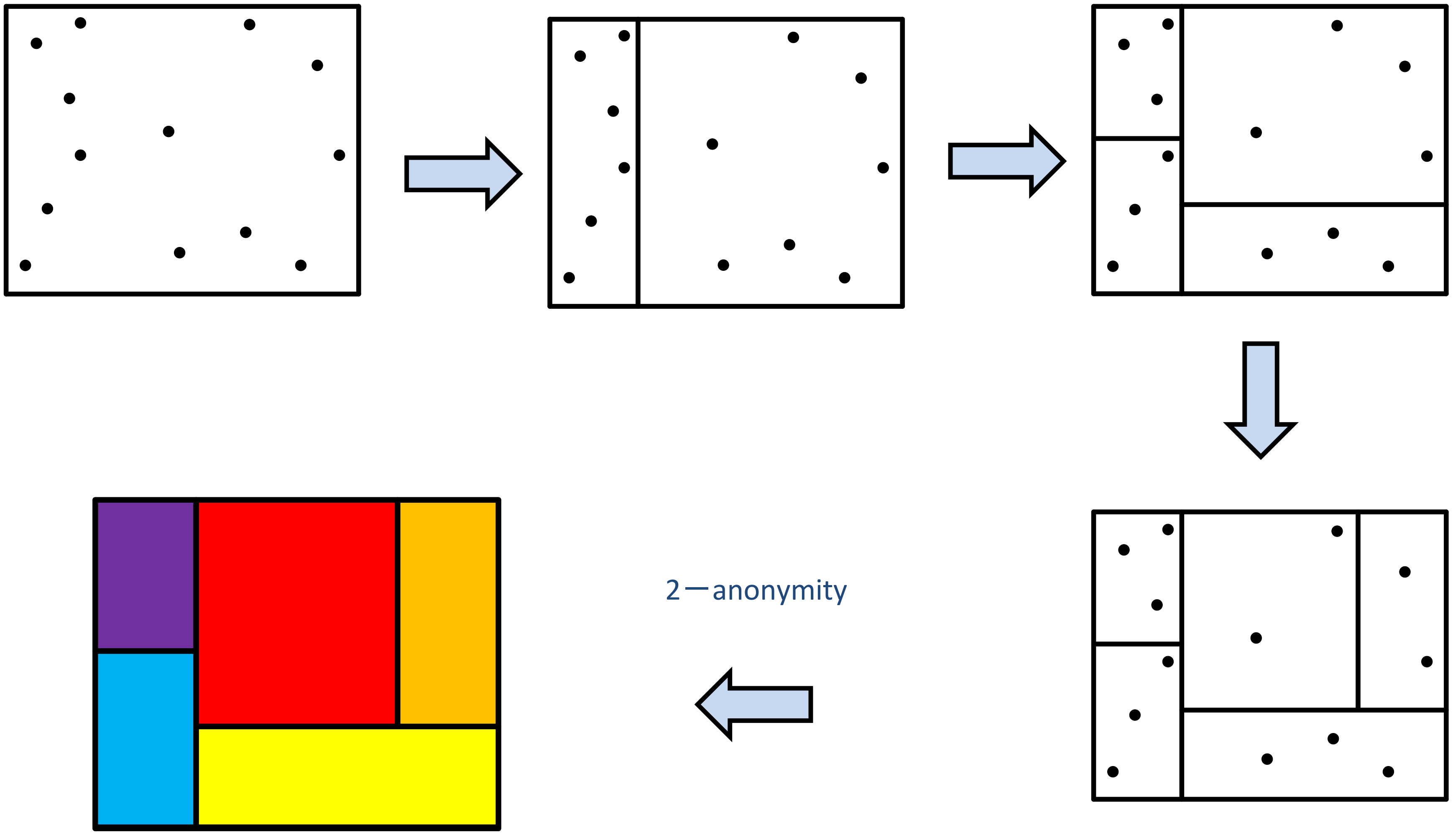
All dimensions are generalized at the same time
topdown [XWP+06]



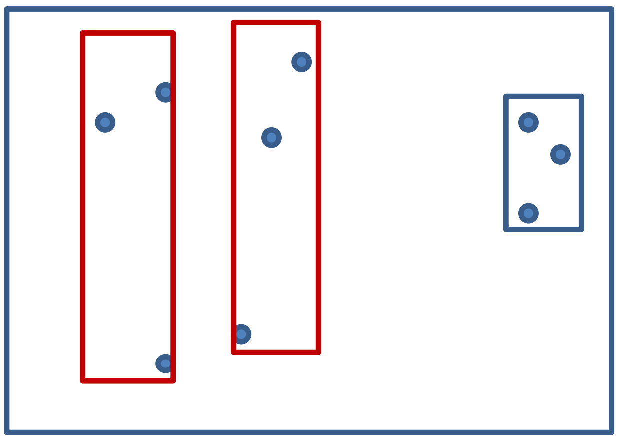


Mondrian

[LDR06]



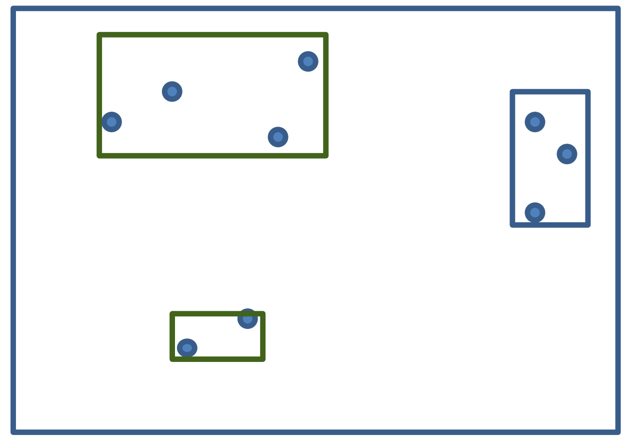
Grouping by boundary length[XWP+06]:



Bad generalization

Long rectangle

Low datamining accuracy



Good generalization

Rectangle near square

High datamining accuracy

Topdown [XWP+06]

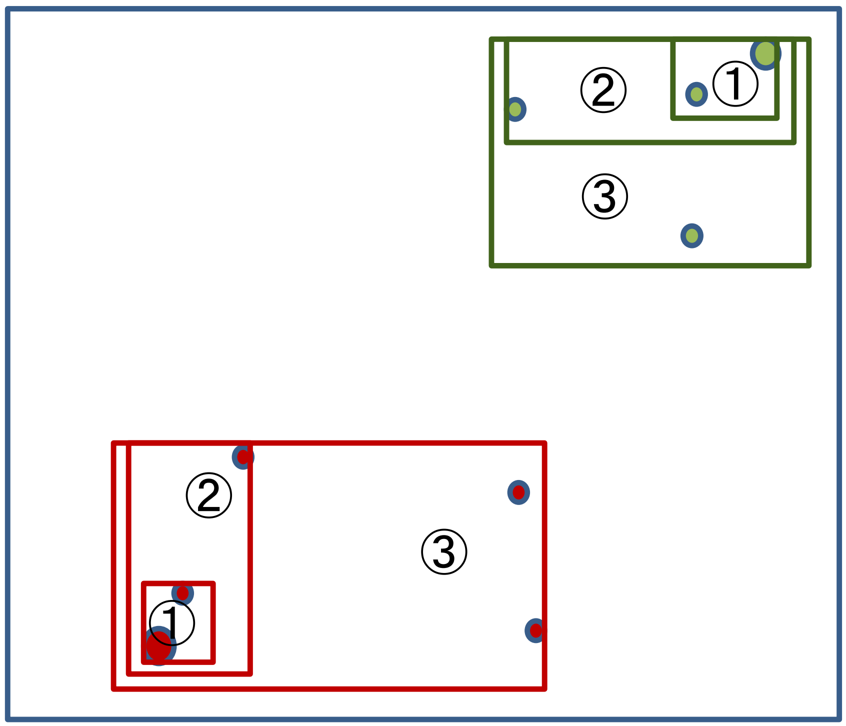
split algorithm

Start with the most distant two data points

- Heuristics
- aggregate to 2 groups from seeds to

The near point is to combined to the group so that the boundary length of the combined group is the minimum among cases other point is combined.

The right figure shows the growing of red and green group by adding ①, ② and ③.



The problem of k-anonymity

- 4-anonymity example
- Homogeneity attack: The third group only consists of cancer patients. Then if combine other DB, the four people in the third group are known to be cancer patients.
- Background knowledge attack: If it is known that in the first group is there one Japanese who has rarely cardiac disease, the Japanese person's illness is inferred as infectious disease.

Anonymous DB

id	Zipcode	age	nationality	disease
1	13053	28	Russia	Cardiac disease
2	13068	29	US	Cardiac disease
3	13068	21	Japan	Infectious dis.
4	13053	23	US	Infectious dis.
5	14853	50	India	Cancer
6	14853	55	Russia	Cardiac disease
7	14850	47	US	Infectious dis.
8	14850	49	US	Infectious dis.
9	13053	31	US	Cancer
10	13053	37	India	Cancer
11	13068	36	Japan	Cancer
12	13068	35	US	Cancer

4-anonymity DB

id	Zipcode	age	nationality	disease
1	130**	<30	*	Cardiac disease
2	130**	<30	*	Cardiac disease
3	130**	<30	*	Infectious dis.
4	130**	<30	*	Infectious dis.
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Cardiac disease
7	1485*	≥40	*	Infectious dis.
8	1485*	≥40	*	Infectious dis.
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

l-diversity

[MGK+06]

- The purpose is that the sensitive information in each group is not skewed.
  - Prevent homogeneity attack
  - Prevent background knowledge attack

l-diversity (intuitive definition)

That a group is l-diverse is defined as at least l kinds of values in the group.

l-diversity algorithm part1

- DB is divided according to each value of sensitive information(disease name).

name	age	sex	disease
John	65	M	flu
Jack	30	M	gastritis
Alice	43	F	pneumonia
Bill	50	M	flu
Pat	70	F	pneumonia
Peter	32	M	flu
Joan	60	F	flu
Ivan	55	M	pneumonia
Chris	40	F	rhinitis

Divide into disease based sub Databases

john	**flu**
Peter	flu
Joan	flu
Bill	flu

Alice	pneumonia
Pat	pneumonia
Ivan	pneumonia

Jack	gastritis

Chris	rhinitis

l-diversity algorithm part2

•Select records from each of left hand side date group and sequentially add each of the right hand side data group. Right hand side record can include Quasi ID of k-anonymity.

Each of these two groups contains at least 3 diseases: 3-diversity

John	**flu**
Peter	flu
Joan	flu
Bill	flu

Alice	pneumonia
Pat	pneumonia
Ivan	pneumonia

Chris	rhinitis

Jack	gastritis

John	**flu**
Joan	flu
Alice	pneumonia
Ivan	pneumonia
Chris	rhinitis

Peter	**flu**
Bill	flu
Pat	pneumonia
Jack	gastritis

Anatomy [Xiaokui06]

- Divide the original table(appeared in l-divesity algorithm part 1) into two tables. The left and right table are only linked by group ID, here 1 and 2.
- 3-diversity

name	age	sex	Group ID
John	65	M	1
Jack	30	M	1
Alice	43	F	1
Bill	50	M	1
Pat	70	F	1
Peter	32	M	2
Joan	60	F	2
Ivan	55	M	2
Chris	40	F	2

Group ID	disease	frequency
1	flu	2
1	pneumonia	2
1	rhinitis	1
2	flu	2
2	pneumonia	1
2	gastritis	1

Data mining is done on these two tables. Since each value is not generalized, the expected accuracy is high.

Side effects of k-anonymity

Defamation

name	age	sex	address	Location at 2016/6/6 12:00
John	35	M	Bunkyo hongo 11	K consumer finance shop
Dan	30	M	Bunkyo Yusima 22	T University
Jack	33	M	Bunkyo Yayoi 33	T University
Bill	39	M	Bunkyo Nezu 44	Y hospital

4-anonymize

name	age	sex	address	Location at 2016/6/6 12:00
John	30’s	M	Bunkyo	K consumer finance shop
Dan	30’s	M	Bunkyo	T University
Jack	30’s	M	Bunkyo	T University
Bill	30’s	M	Bunkyo	Y hospital

Dan , Jack and Bill are not recognized a person different from John by 4-anonyumity, all four persons are suspected to stay at K consumer finance shop→k-anonymization provokes defamation on Dan, Jack and Bill.

k-anonymity provokes defamation in sub area aggregation

l-diversity makes situation worse

name	age	sex	address	Location at 2016/6/6 12:00
John	35	M	Bunkyo hongo 11	K consumer finance shop
Dan	30	M	Bunkyo Yusima 22	K consumer finance shop
Jack	33	M	Bunkyo Yayoi 33	K consumer finance shop
Bill	39	M	Bunkyo Nezu 44	K consumer finance shop

These values shows all four is at K consumer finance shop

Exchange one person to make DB 2-diversity

name	age	sex	address	Location at 2016/6/6 12:00
John	30's	M	Bunkyo	K consumer finance shop
Dan	30's	M	Bunkyo	K consumer finance shop
Jack	30's	M	Bunkyo	K consumer finance shop
Alex	30's	M	Bunkyo	T Univeristy

By 2-diversifying, Ales becomes strongly suspected to be at K consumer finance shop → *l*-diversity provokes defamation

Why defamation happens?

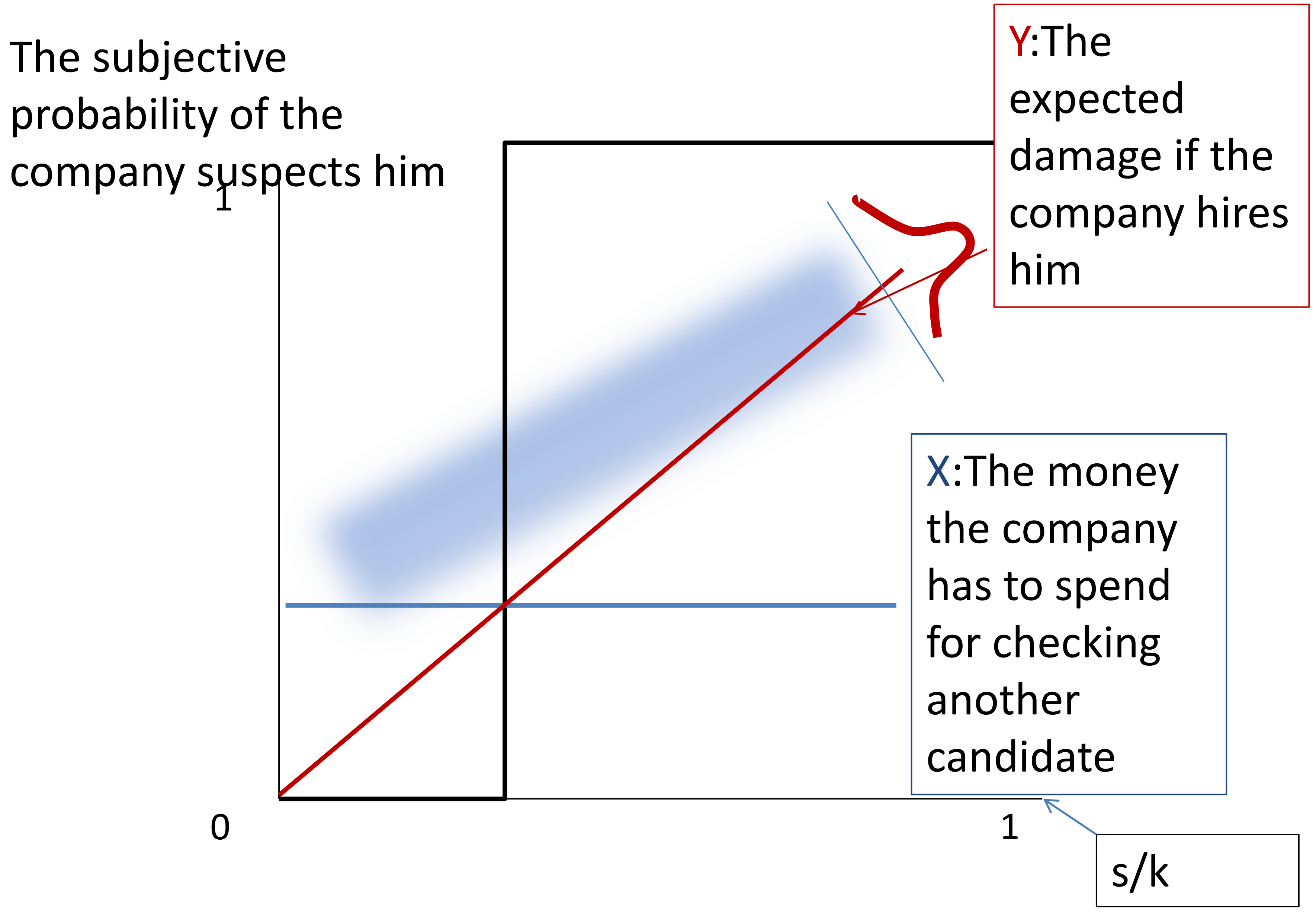
- Case study
  - A job candidate who is a good university student.
  - He is in k people group that includes at least one person who went to a consumer finance shop.
  - A company he tries to take entrance examination does not want hire a person who goes to a consumer finance shop.
  - He is suspected to go to a consumer finance shop.→ defamation!

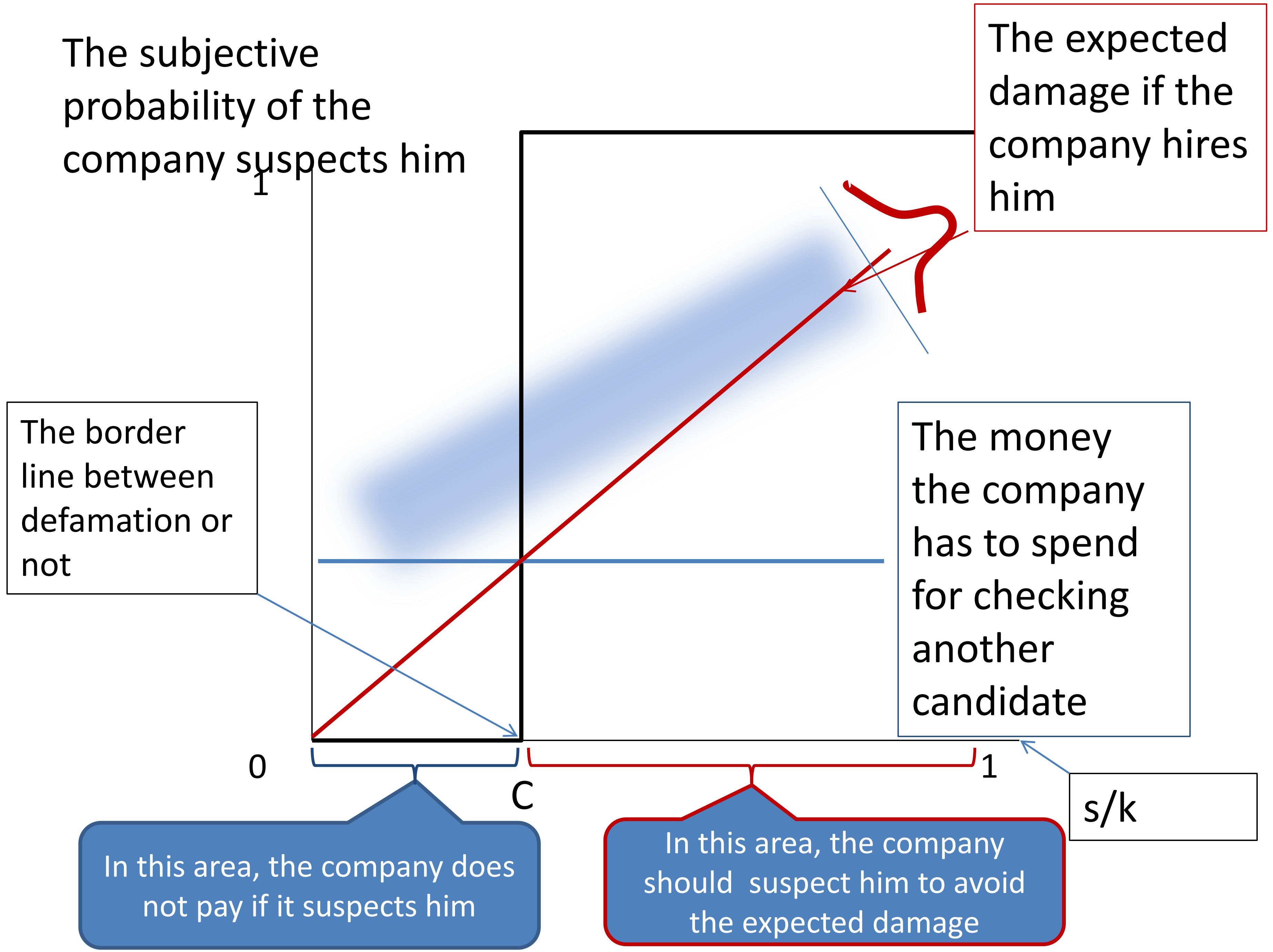
Back ground situation of defamation

- Case study cont.
  - If the company deletes him from candidates, it must use another time and money, say X, to check another candidate:
  - If the company hires a bad buy, it will suffer a certain amount of damage, say Y, by his bad behavior.
  - Then if the expected value of Y is more than X, the company becomes very negative, otherwise not negative about him.
  - This is a defamation model from an economical point of view.

Back ground situation of defamation

- Case study cont.
  - Another factor is the probability that he actually went to a consumer finance shop.
  - This probability is proportional to the number of consumer finance shop visitors, say s, in k people of k-anonymity group = s/k.
  - Y is proportional to s/k
  - Then the relation is sketched in the figure on next slide.


The subjective probability of the company suspects him
1
0
1
Y:The expected damage if the company hires him
X:The money the company has to spend for checking another candidate
s/k


The subjective probability of the company suspects him
1
The expected damage if the company hires him
The border line between defamation or not
The money the company has to spend for checking another candidate
0
C
1
s/k
In this area, the company does not pay if it suspects him
In this area, the company should suspect him to avoid the expected damage

Solution

- Then the solution is simple:
  - Make the border line as small as possible.
  - But how?

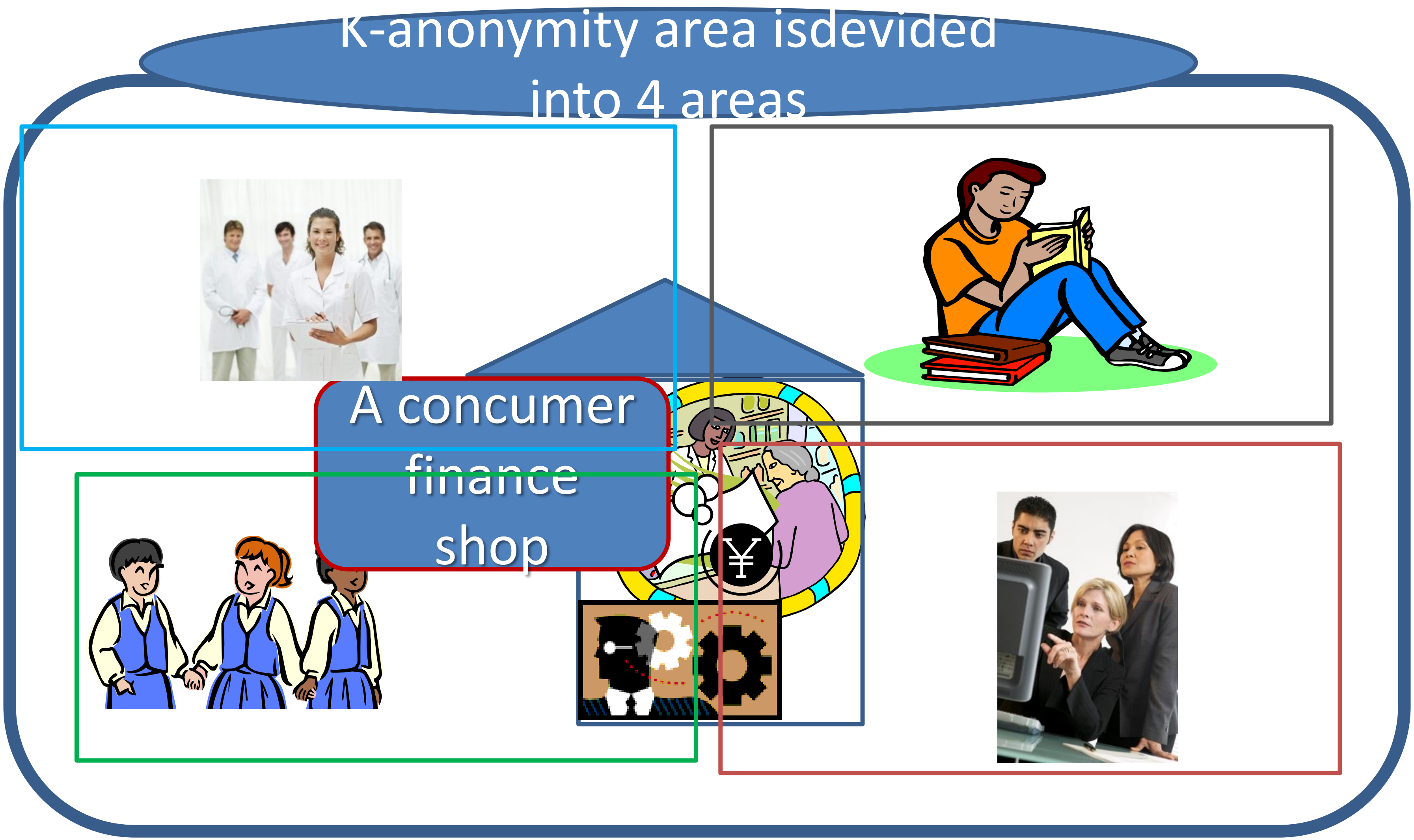
- We can revise k-anonymization algorithm in order to minimize the number of bad behavior guys in k-anonymity group.
  - This revision, however, reduce the accuracy of the data.
  - Then the problem comes to be a optimization problem:

Maximize   Accuracy of data

subject to number of bad guys ≤ 1

$\text{in k-anonymity group}$

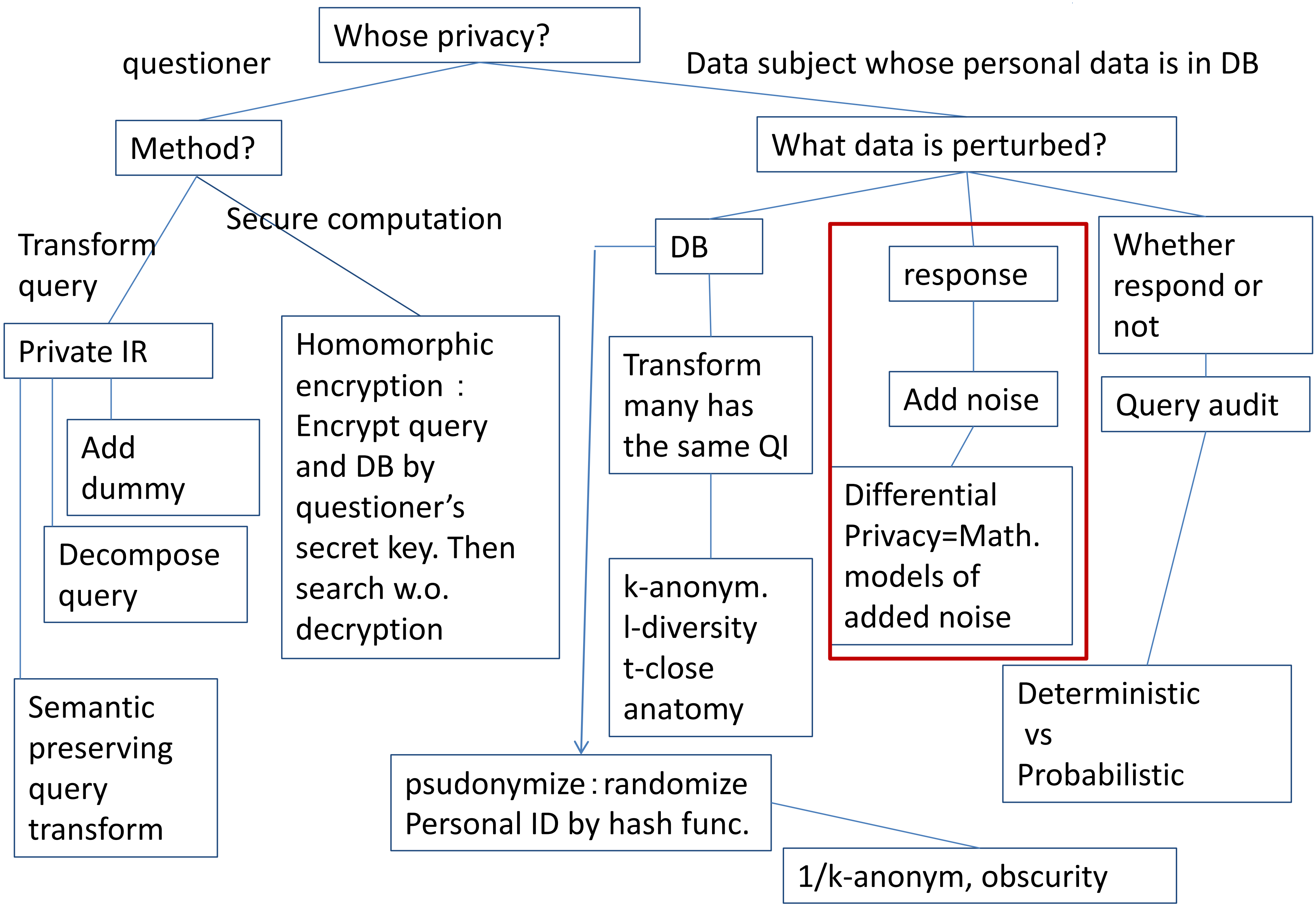
A consumer finance shop is devided into 4 parts to reduce # of poepole visit it is less or equal than one



Outline of proposed algorithm

1. Do k-anonymization.
2. If one group includes more than one bad guys
   ① Then combine this and two nearest groups
   ② Do k-anonymization to this combined group to make two groups that includes at most one bad guys.
   ③ If step ② fails,
   ④ then go back to one step in 1. Do k-anonymization, namely try to find another generalization in k-anonymization.

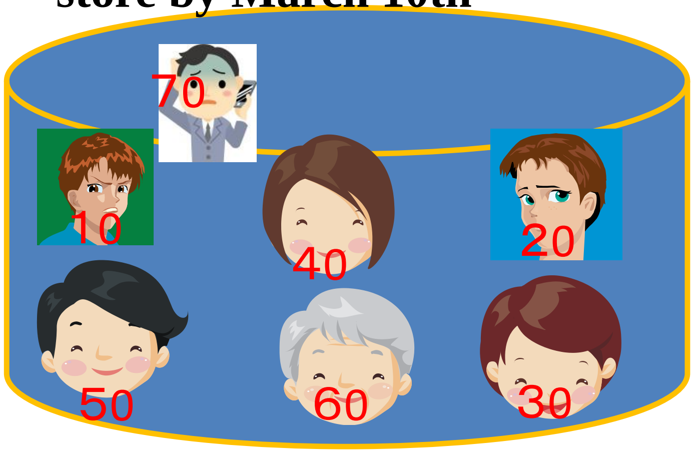
Overview of Privacy Protection Technologies



Differential Privacy: DP

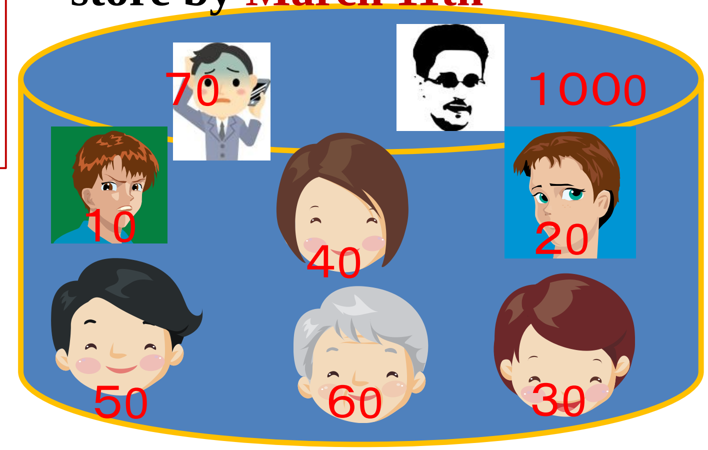
Motivation of DP

DB : ***D*(sales data of jewel store by March 10th**



He is known to come to the store on March 11

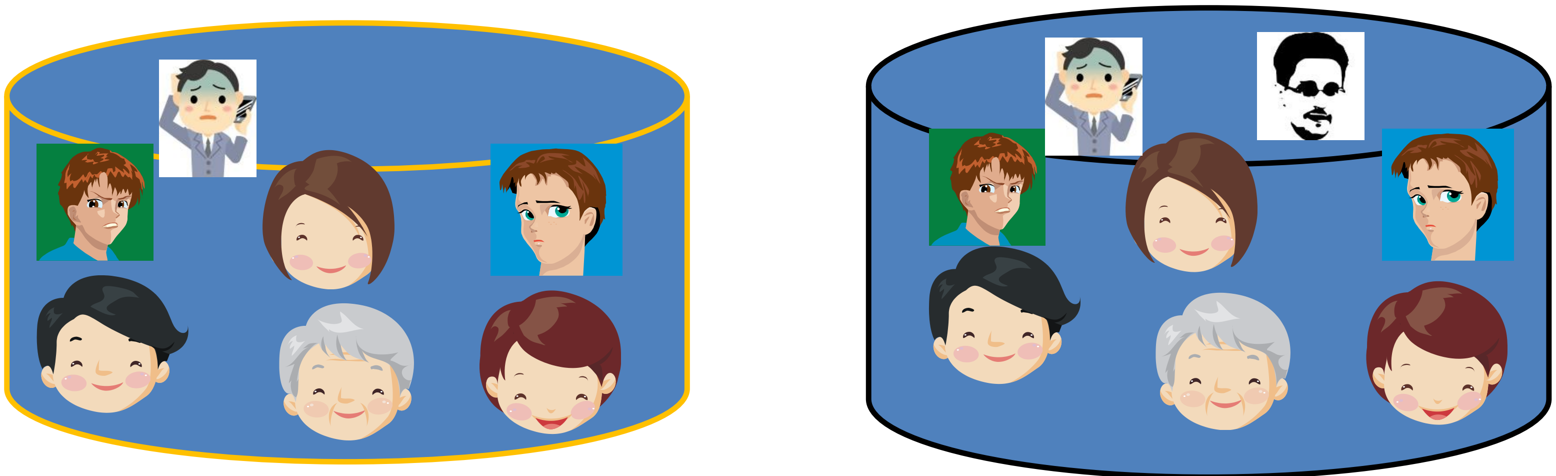
DB : ***D*(sales data of jewel store by March 11th**



- A query is the highest price (red number) paid by customers.
- The highest till March 10th is 60,000 yen. It becomes 1,000,000 yen on March 11th.
- If some one sees  in the store and gets the answer of 10 th and 11th, he/she gets the information about  that is he bought a jewel of 1,000,000 yen , and probably is very rich.
- This privacy breach is avoided if we add some noise to the answer: → DP

Simple Example

DB : ***D*** DB : ***D’***



D differs from *D’* only by one record of  .

We want to prevent a questioner from realizing which DB, say D or D’ is used to make an answer. For this purpose, DP adds a noise to the answer.

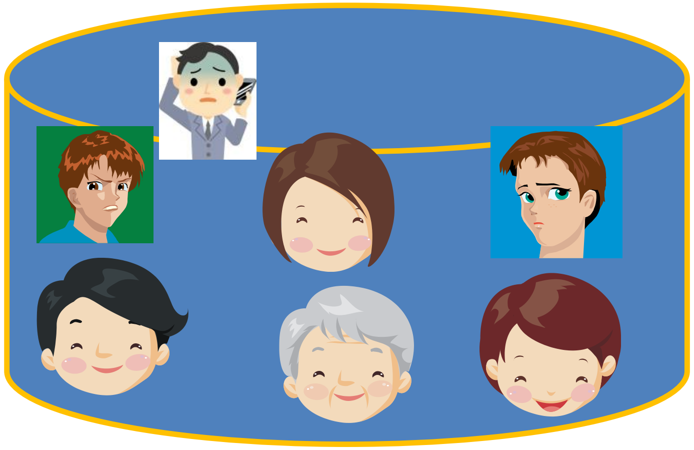
- example： A question is the number of men and women in DB.

    - If no noise is added, the answer from *D* is 4 men and 3 women,
    - the answer from *D’* is 5 men and 3 women.

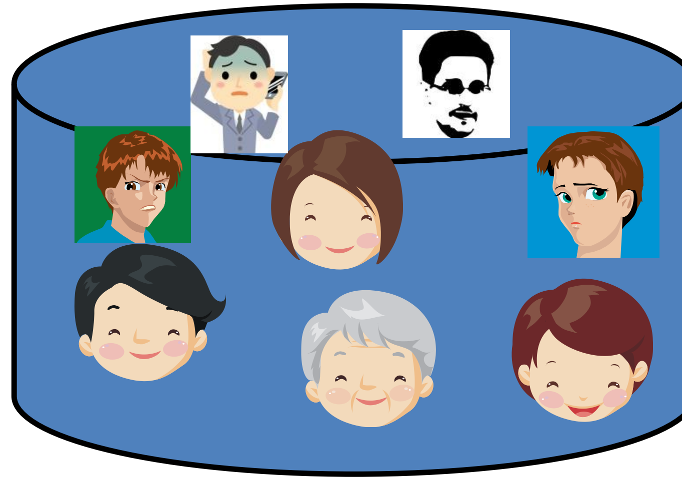
    - Then *D’* is known to have one more man than *D*.
    - → There is a chance to realize that  is in *D’*.

Simple Example cont.

DB : D



DB : D'



- Then D' is known to have one more man than D.
- → There is a chance to realize that  is in D'.

- DP adds a noise as follows: Add 1 to the answer of men number of D.
- Add -1 to the answer of men number of D'.

- Then , the answer from D is（5 men， 3 women）、that from D' is（4 men， 3 women） → The questioner does not know whether  is in DB or not.

- It is a strong privacy protection if the existence it self is concealed .

How large a noise should be?

- In the above figure, X00 means that a year income is X,000,000 yen.
- The highest income in *D* is 8,000,000yen, and that of *D'* is 15,000,000 yen.

- A question of the highest year income reveals that *D'* includes a high income person. 

- In order to prevent this breach, we should add something like 7,000,000 yen = 15,000,000-8,000,000 yen. It is so big that the accuracy or usefulness of DB is impaired very
- More accurately, a size of noise should be heavily related to the largest difference of answer from *D* and that of *D'*.
- This largest difference is called sensitivity in DP.

DP is

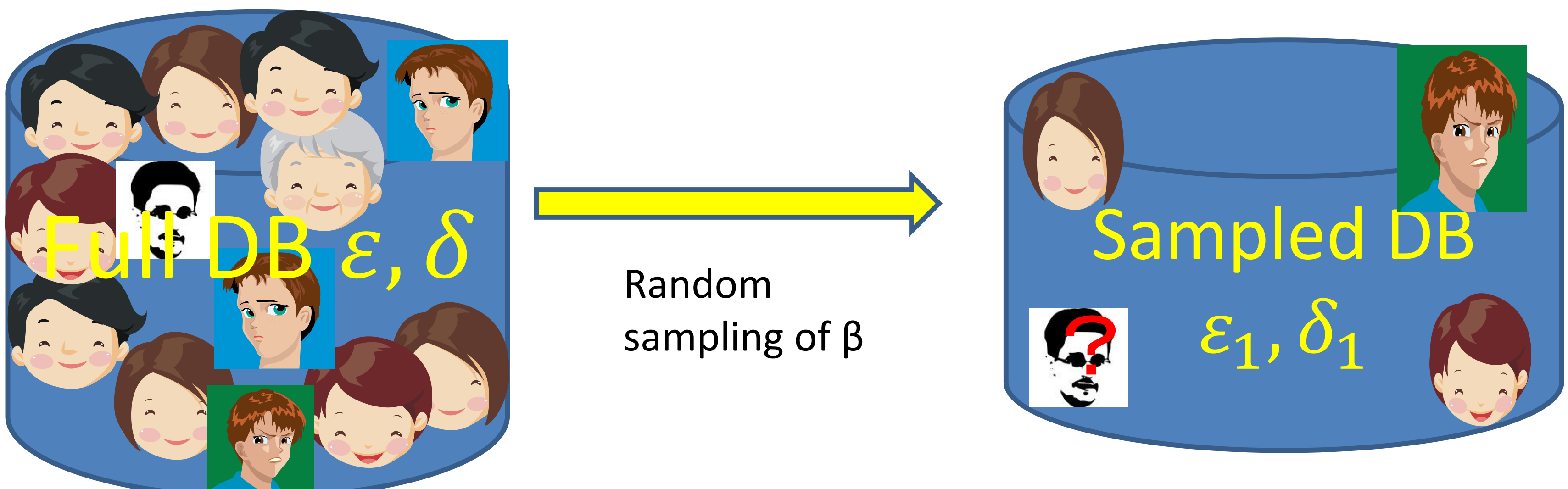
- For the most similar pair of DBs, say, only one record is different, D and D'
- A query is asking a function such as a sum of the specified attribute,

- Then DP is a mechanism of adding a certain noise to the answer in order not to be recognized which of DBs are used . *f*(*D*) (or *f*(*D'*)) is a noise added answer

- $(\varepsilon, \delta) - DP$ is the following
- $\forall D, D' \;\; P\big(f(D)\big) \leq e^{\varepsilon} P\big(f(D')\big) + \delta$

Randomly sampled DB

- The purpose of DP is not to be recognized the existence of 
- In a sampled DB, to decide whethe  is in the DB is difficult
- When sampling rate is β, then the noise ε_1, δ_1 to add is smaller than the full DB case ε , δ .
- $\boxed{\delta_1 = \beta\delta, \quad e^{\varepsilon_1} - 1 = \beta(e^{\varepsilon} - 1)}$



Time is too short to tell the whole technology detail.

If you would like to know more, please read this book.

Reference

- [LDR 05]LeFevre, K., DeWitt, D.J., Ramakrishnan, R. Incognito: Efficient Full-domain *k*-Anonymity. SIGMOD, 2005.
- [LDR06]LeFevre, K., DeWitt, D.J., Ramakrishnan, R. Mondrian Multidimensional *k*-Anonymity. ICDE, 2006.
- [XWP+06] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A., Utility-Based Anonymization Using Local Recoding. SIGKDD, 2006.
- [MGK2007]MACHANAVAJJHALA,A. KIFER,D. GEHRKE,J. and VENKITASUBRAMANIAM, U. l-Diversity: Privacy Beyond *k-Anonymity.* ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, Article 3,2007
- [S01] Samarati, P. Protecting Respondents' Identities in Microdata Release. IEEE TKDE, 13(6):1010-1027, 2001.
- [S02a] Sweeney, L. *k*-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.
- [S02b] Sweeney, L. *k*-Anonymity: Achieving *k*-Anonymity Privacy Protection using Generalization and Suppresion. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.
- Ninghui Li,Tiancheng Li,Venkatasubramanian, S. “t-Closeness: Privacy Beyond k-Anonymity and –Diversity”. ICDE2007, pp.106-115, 2007.
- [SMP] Sacharidis, D., Mouratidis, K., Papadias, D. *k*-Anonymity in the Presence of External Databases（to be appeared)
- [Xiaokui06] X. Xiaokui and T. Yufei. (2006). Anatomy: Simple and Effective Privacy Preservation. VLDB, 139-150.
- [Dwork & Roth] Dwork C. and A.Roth. (2013). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science Vol.9 Nos. 3-4, 211-407 DOI: 10.1561/0400000042.
- [Li,Qardaji,Su2012] Ninghui Li, Wahbeh Qardaji, Dong Su: On Sampling, Anonymization, and Differential Privacy: Or, *k*-Anonymization Meets Differential Privacy. Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security(ASIACCS’12). Pages 32-33. 2012