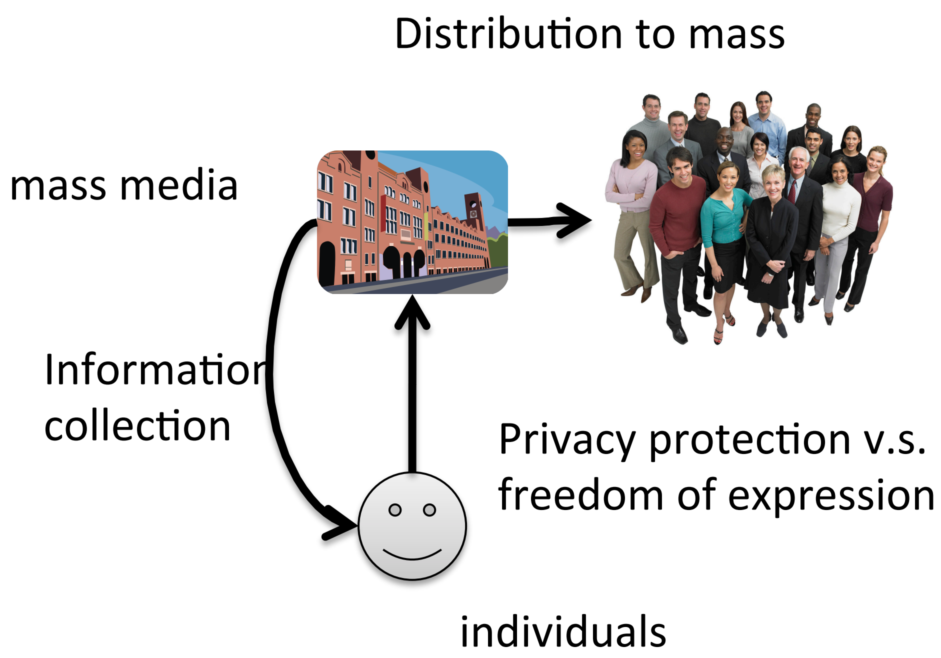
PHUSE: Talk
Privacy-Enhancing Technologies for Personal Data Analysis

Jun Sakuma

U. Tsukuba CS dept. / JST CREST

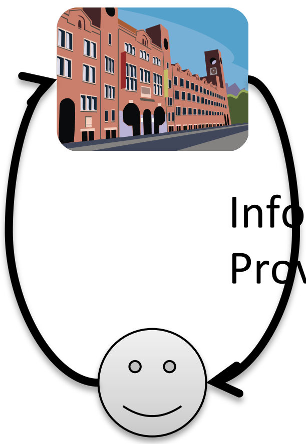
What is privacy?

- Privacy right (classical)
  - The right to be let alone
  - Only publishers, newspaper, and broadcast stations could distribute information to mass

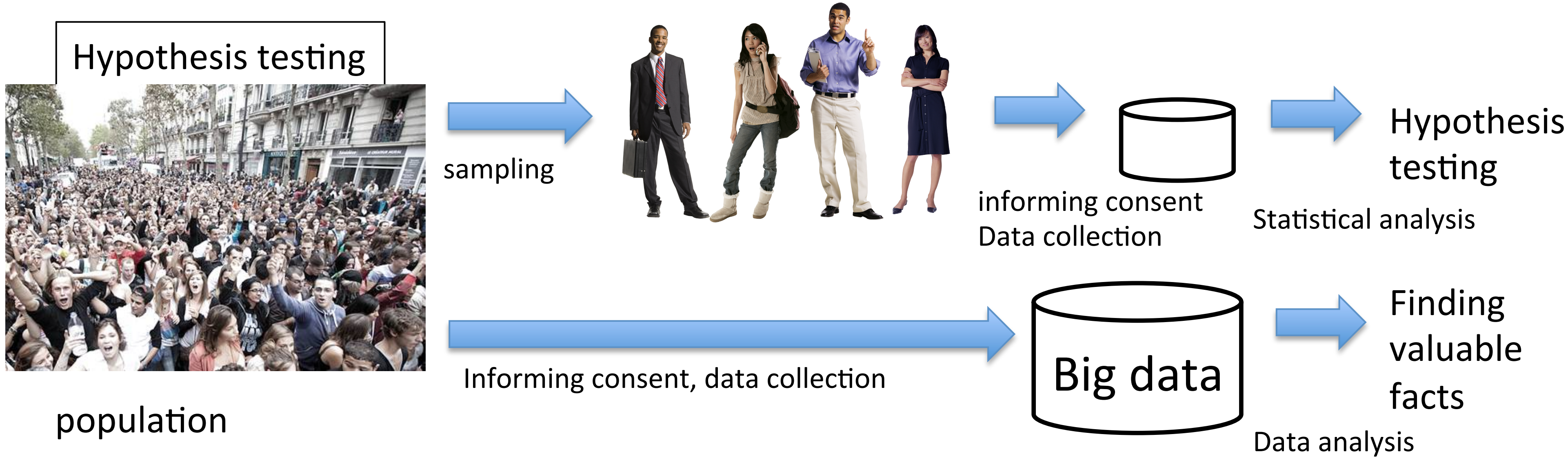


What is privacy?

- Claim of privacy right
  - Rights to control (correction, removal, etc.) my information managed by others
  - OECD ”Privacy Principles” Collection Limitation Principle, Data Quality Principle, Purpose Specification Principle, Use Limitation Principle, Security Safeguards Principle, Openness Principle, Individual Participation Principle, Accountability Principle
- Assumes personal information is collected by public organizations or private companies



Big data analysis



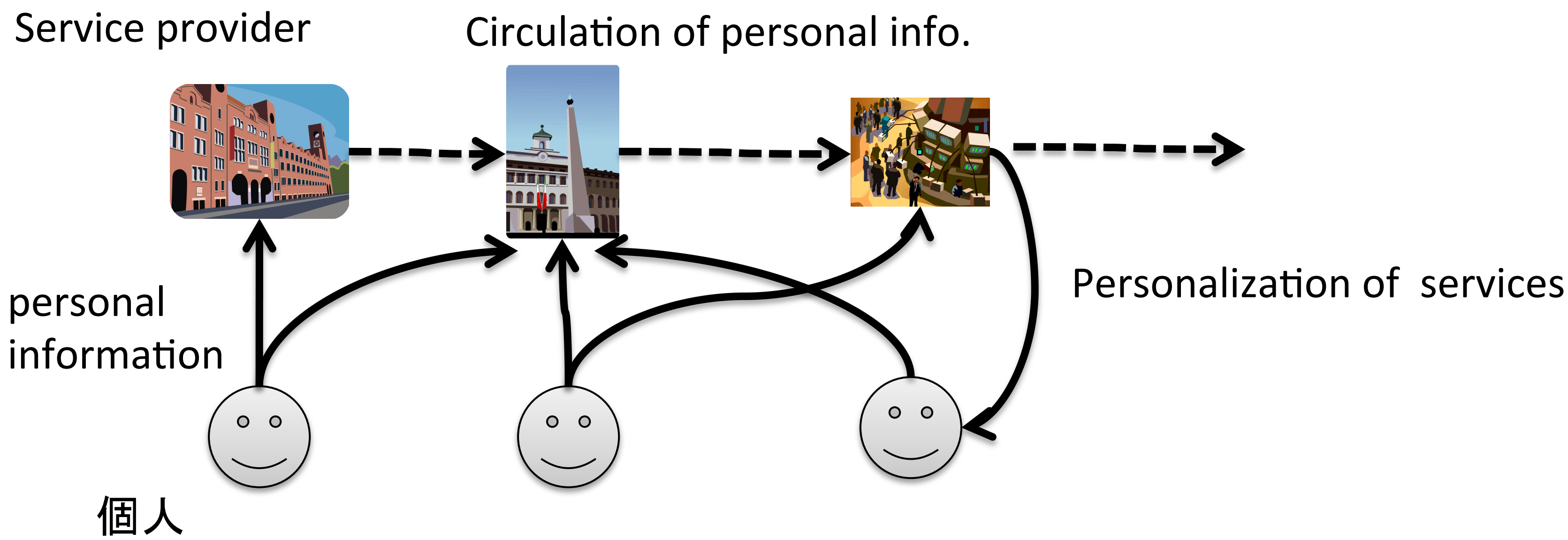
Federation of big data increases service opportunities exponentially

	data	utilization
transportation	Travel log	Traffic prediction
retail	Purchase history	Item recommendation
hospital	EHR	medication
sensors	Vital data	Health management
…	…	…

Federation of big data increases service opportunities exponentially

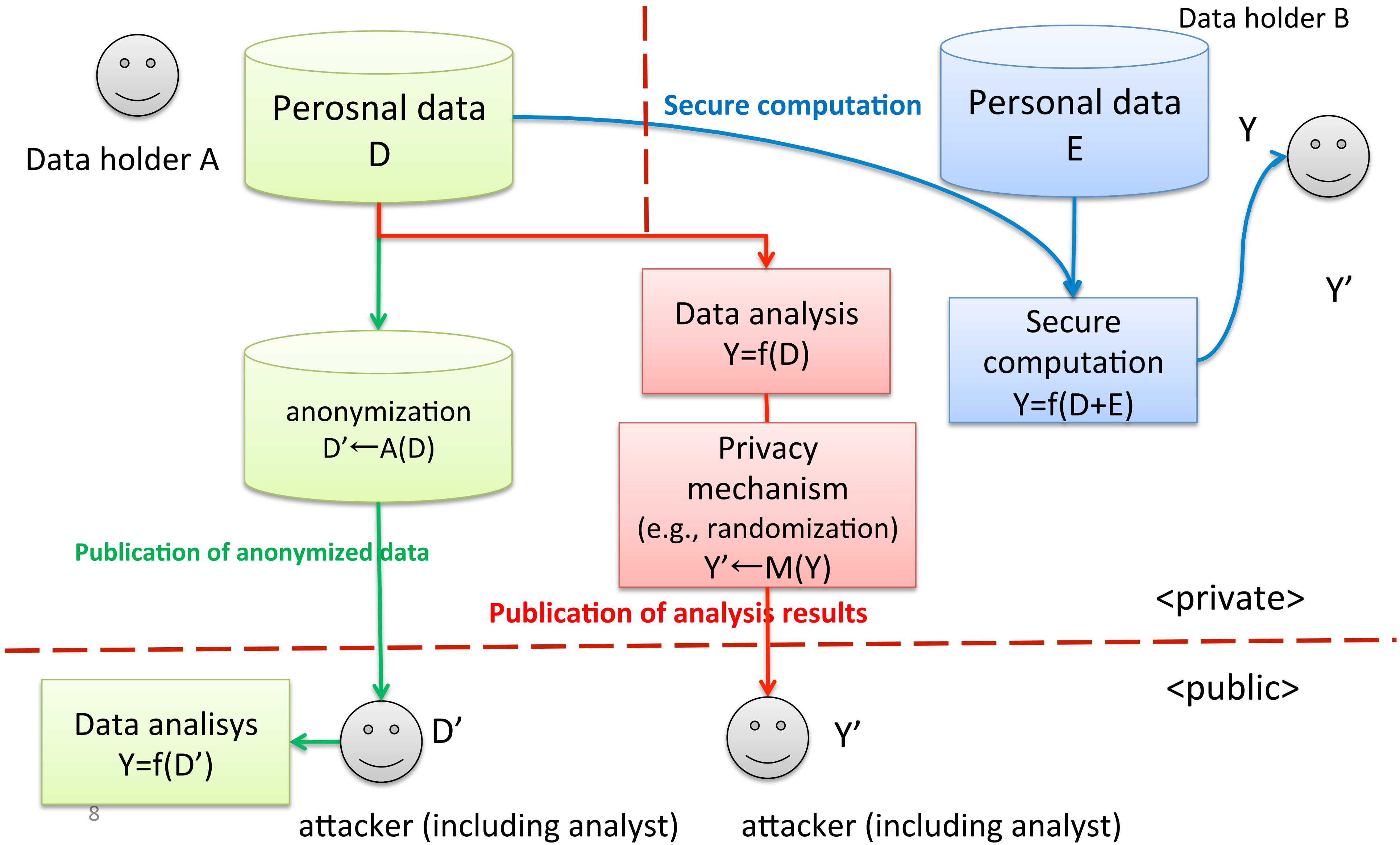
		transportation	retailers	hospitals	sensors	…
		Travel log	Purchase history	EHR	vital	…
transportation	Travel log	Traffic prediction	NEW! Location-aware Advertisement	NEW! Health management with	NEW! Stress measurement and people flow	
retailers	Purchase history		Item recommendation	NEW! Common diseases and dietary habit	NEW! Shopping behavior and vital signs	
hospital	EHR			medication	NEW! Safety confirmation at home	
sensors	Vital data				Health management	
…	…	…				…

Personal information in big data era



- Circulation of High-reso. and high-freq personal info
- Federation of personal info from multiple sources
- Privacy protection

Models of data analysis with personal info.



Privacy enhancing technologies

- Information security
  - Computational theory, cryptography, system engineering, network engineering
- Data privacy
  - Cryptography, information theory, statistics, database, data mining, machine learning, mechanism design, etc.
  - For operation of services: low and legal systems, ethics, phycology, information economy, etc.
- Huge amount of collective expertise from diverse domains

Components of personal information

- Direct identifiers:
  - identify an individual by itself
  - My number (Japan), SSN (US), driver's license no.,
  - customer ID (for those who hold customer ID-Name mapping)
  - Name, phone number, physical address, email address
- Quasi-identifiers:
  - identify an individual by combination
  - Age, gender, zip-code, race, hair color, occupation, etc.
- Equivalence class

ID	Sex	Age	Profession	Drug Test
1	Male	37	Doctor	Negative
2	Female	39	Doctor	Positive
3	Male	37	Doctor	Negative
4	Male	39	Doctor	Positive
5	Male	39	Doctor	Negative
6	Male	37	Doctor	Negative

Components of personal information

- Sensitive data: Information that could harmful if identified
  - Time independent :
    - party affiliation, religious affiliation, past medical history, etc.
  - Time dependent :
    - Travel history, bank account history, purchase history, etc…
- Sensitive data is context dependent
  - Past medical history: broken bone, diabetes, and HIV
    - What if a scout finds “broken bone” of a promising candidate?
    - What if health insurance company finds “unreported diabetes” of customers?

Running example

Categorical Att.

Numerical Att.

A value of the age att.

attribute

record

Patient ID	Name	Address	age	hospital	disease	medicine
0A99934	Tsukuba Taro	Ibaraki, Tsukuba city, XX	32	Hospital A	backache	XXX
1B92372	Toukei Hanako	Tokyo, Machida city, YY	46	Clinic B	Hay fever	YYY
3C88892	Tokumei Yukie	Kyoto, Uji city, ZZ	42	Hospital C	Lung cancer	ZZZ
…	…	…	…	…	…	…

Direct identifier

Quasi identifier

Sensitive data

Privacy risk 1: identify inference (de-identification)

Name	Address	age	hospital	disease	medicine	Name
Tsukuba Taro	Ibaraki, Tsukuba city, XX	32	Hospital A	backache	XXX	Tsukuba Taro
Toukei hanako	Tokyo, Machida city, YY	46	Clinic B	Hay fever	YYY	Toukei hanako
Tokumei Yukie	Kyoto, Uji city, ZZ	42	Hospital C	Lung cancer	ZZZ	Tokumei Yukie
…		…	…		…	…

One who visits Clinic, lives in Machida city in Tokyo, and is 46 yeras-old is “Tokumei Hanako”.

Privacy risk 2: attribute inference

Name	address	age	hospital	disease	Name
Tsukuba Taro	Ibaraki, Tsukuba city, XX	32	Hospital A	backache	XXX
Toukei hanako	Tokyo, Machida city, YY	46	Clinic B	Hay fever	YYY
Tokumei Yukie	Kyoto, Uji city, ZZ	42	Hospital C	Lung cancer	ZZZ
…	…	…		…	…

One who visits Clinic, lives in Machida city in Tokyo, and is 46 years-old has “Hay fever”.

Privacy risk 3: linkage inference

Disclosed database A

name	disease	address	age	Aff.
Tsukuba Taro	backache	Ibaraki	32	University A
Toukei hanako	Hay fever	Tokyo	46	Company B
Tokumei Yukie	Lung cancer	Kyoto	42	Company C
…	…	…	…	…

Disclosed database B

name	disease	address	age	Aff.	name
Tsukuba Taro	backache	Ibaraki	32	Hospital A	XXX
Toukei hanako	Hay fever	Tokyo	46	Clinic B	YYY
Tokumei Yukie	Lung cancer	Kyoto	42	Hospital C	ZZZ
…	…	…	…	…	…

Adversary learns linkage between records

name	disease	address	age	Aff.	Hos	name
Tsukuba Taro	backache	Ibaraki	32	University A	Hospital A	XXX
Toukei hanako	Hay fever	Tokyo	46	Company B	Clinic B	YYY
Tokumei Yukie	Lung cancer	Kyoto	42	Company C	Hospital C	ZZZ
…	…	…	…	…	…	…

One who works for company B and lives Tokyo visits clinic B and has hay fever

Statistical disclosure control

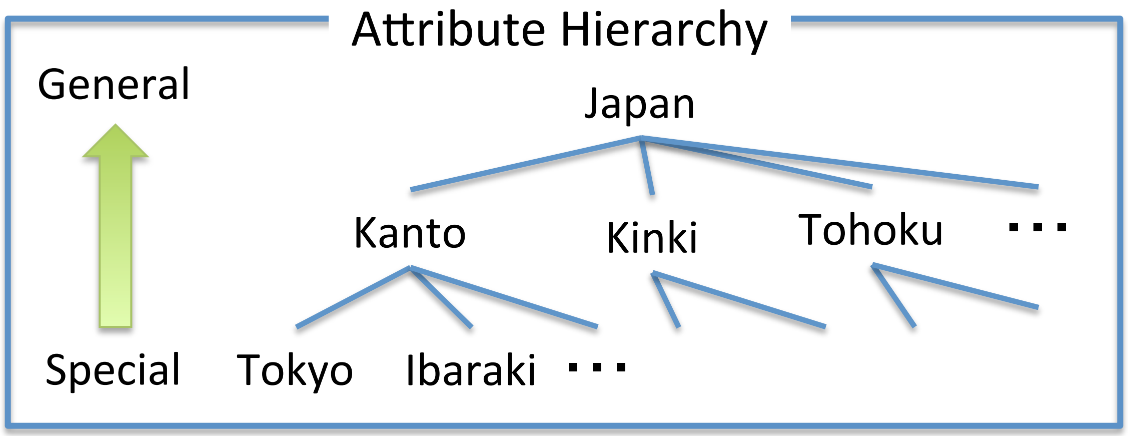
- Top(bottom)-coding
  - Ex1) Over 100 (if the age is more than 100)
  - Ex2) Under 10 (if the age is less than 10)

Address	Age	Hospital	Disease	Medicine
Ibaraki, A city	32	Hospital A	Backache	AAA
Tokyo, B city	46	Clinic B	Rhinitis	BBB
Kyoto, C city	42	Hospital C	Cancer	CCC
Niigata, D city	102	Hospital D	Pneumonia	DDD
Hokkaido, E city	3	Clinic E	Otitis media	EEE
…	…	…	…	…

Address	Age	Hospital	Disease	Medicine
Ibaraki, A city	32	Hospital A	Backache	AAA
Tokyo, B city	46	Clinic B	Rhinitis	BBB
Kyoto, C city	42	Hospital C	Cancer	CCC
Niigata, D city	Over 100	Hospital D	Pneumonia	DDD
Hokkaido, E city	Under 10	Clinic E	Otitis media	EEE
…	…	…	…	…

Statistical disclosure control

- Generalization



Address	Age	Hospital	Disease	Medicine
Ibaraki, A city	32	Hospital A	Backache	AAA
Tokyo, B city	46	Clinic B	Rhinitis	BBB
Kyoto, C city	42	Hospital C	Cancer	CCC
Niigata, D city	102	Hospital D	Pneumonia	DDD
Hokkaido, E city	3	Clinic E	Otitis media	EEE
…	…	…	…	…

Address	Age	Hospital	Disease	Medicine
Kanto	32	Hospital A	Backache	AAA
Kanto	46	Clinic B	Rhinitis	BBB
Kinki	42	Hospital C	Cancer	CCC
Chubu	102	Hospital D	Pneumonia	DDD
Hokkaido	3	Clinic E	Otitis media	EEE
…	…	…	…	…

Statistical disclosure control

- Cut off
  - Ex1) Cut off some rare (or critical) diseases.
  - Ex2) Cut off some rare ages.

Address	Age	Hospital	Disease	Medicine
Ibaraki, A city	32	Hospital A	Backache	AAA
Tokyo, B city	46	Clinic B	Rhinitis	BBB
Kyoto, C city	42	Hospital C	Cancer	CCC
Niigata, D city	102	Hospital D	Pneumonia	DDD
Hokkaido, E city	3	Clinic E	Otitis media	EEE
…	…	…	…	…

Address	Age	Hospital	Disease	Medicine
Ibaraki, A city	32	Hospital A	Backache	AAA
Tokyo, B city	46	Clinic B	Rhinitis	BBB
Kyoto, C city	42	Hospital C		CCC
Niigata, D city		Hospital D	Pneumonia	DDD
Hokkaido, E city	3	Clinic E	Otitis media	EEE
…	…	…	…	…

Protection from identity inference

How much do we need to modify the table?

ID	name	address	age	hospital	disease	medicine
0A99934	筑波太郎	Ibaraki, A city	32→30s	A病院	backache	XXX
1B92372	統計花子	Tokyo, B city	46→40s	B医院	Hay fever	YYY
3C88892	匿名幸	Kyoto, C city	42→40s	C病院	HIV→removal	ZZZ
…	…	…	…	…	…	…

(de-identification) Global recoding removal Local suppression

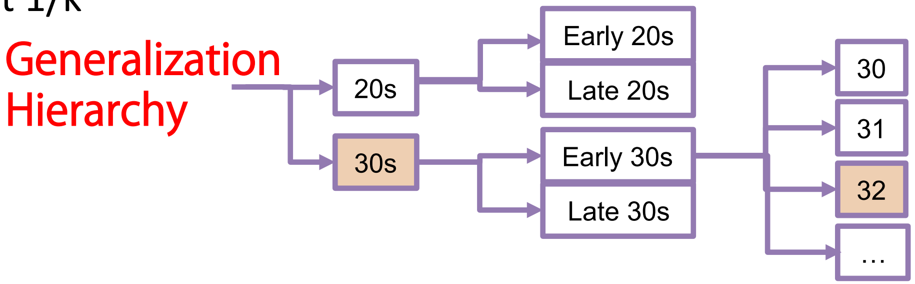
To what extent do we need modification?

Phuse de-identification standard

- Scramble subject IDs (de-identification)
- Anchor the first date and add some modification
- QIs are kept as it is if necessary
- Country is elevated to continent (generalization)
- Birth date is changed to age (<89) (top-coding)
  - categorize by 5-year interval (generalization)
- Free texts are removed

Stronger anonymization: k-anonymity [Sweeney02]

- I know a person in the table who lives in 212-0014 is 24 years-old
  - The person in the table is identifiable (identity inference)
- The size of equivalence class is at least k (= k-anonymity)
- The probability of identity inference is at most 1/k



Original data

Client ID	Name	Post code	Age	Disease
12245678	Tsukuba Taro	233−0011	26	Diabetes
12245679	Toukei Hanko	233−0011	54	Backache

De-identification

De-identification

Client ID	Name	Post code	Age	Disease
		232−0011	26	Hernia
		232−0015	34	Backache
		232−0017	27	Backache
		233−0012	45	Rhinitis
		233−0013	43	Asthma
		233−0014	42	Tuberculosis
		233−0014	23	Diabetes
		233−0014	24	Diabetes
		233−0014	26	Diabetes

Delete

Direct identifier — Quasi identifier — Sensitive data

k-anonymization (resistant to identity inference)

k-anonymization

Post code	Age	Disease
232−001x	20−39	Hernia
232−001x	20−39	Backache
232−001x	20−39	Backache
233−001x	40−49	Rhinitis
233−001x	40−49	Asthma
233−001x	40−49	Tuberculosis
233−0014	20−29	Diabetes
233−0014	20−29	Diabetes
233−0014	20−29	Diabetes

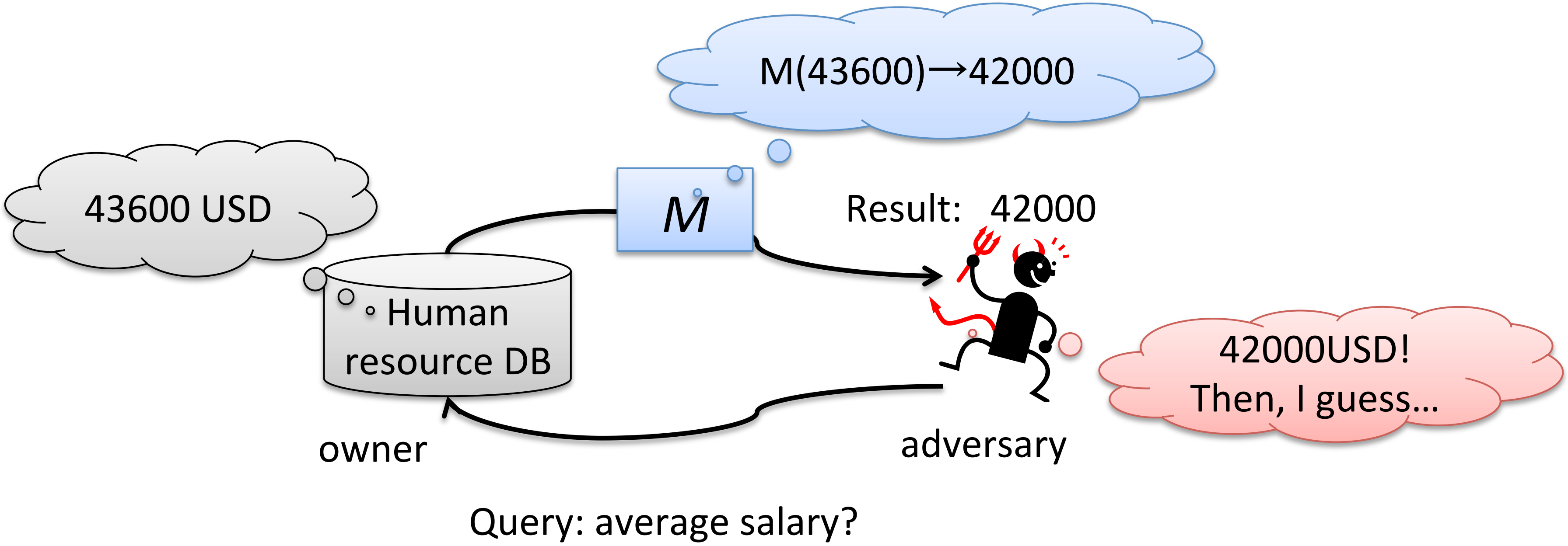
3

3

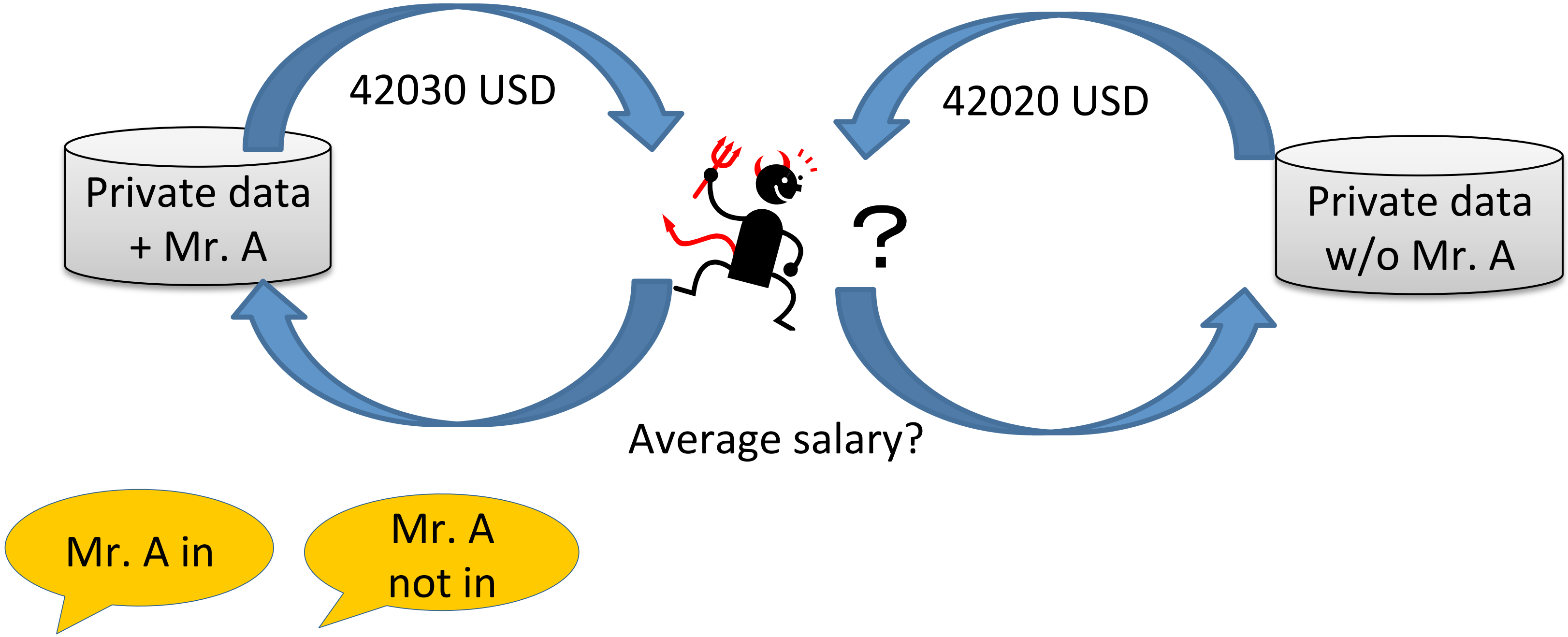
3

SWEENEY, Latanya. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10.05: 557-570.

Output privacy



Semantic secuirty

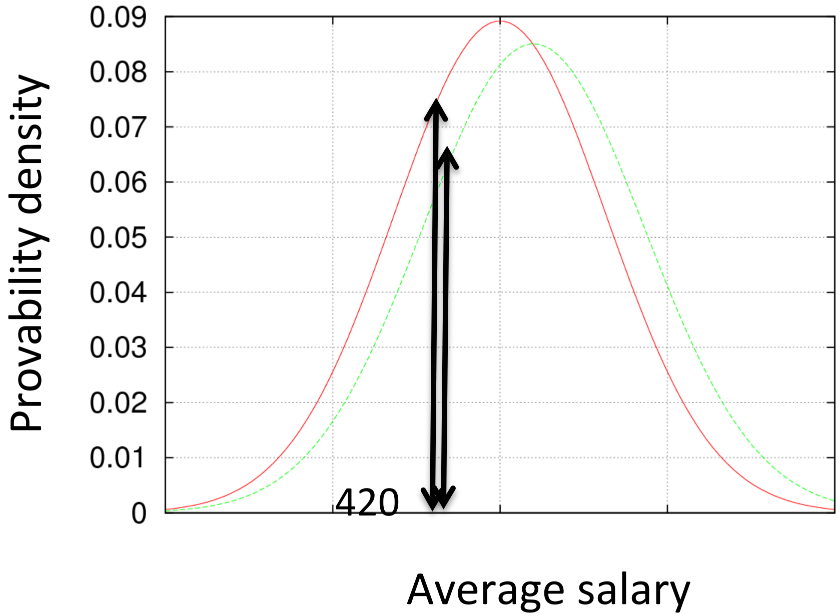


- If f(D+A) and f(D) are not that different, let us regard disclosing f(D+A) is not that privacy invasive

Differential Privacy

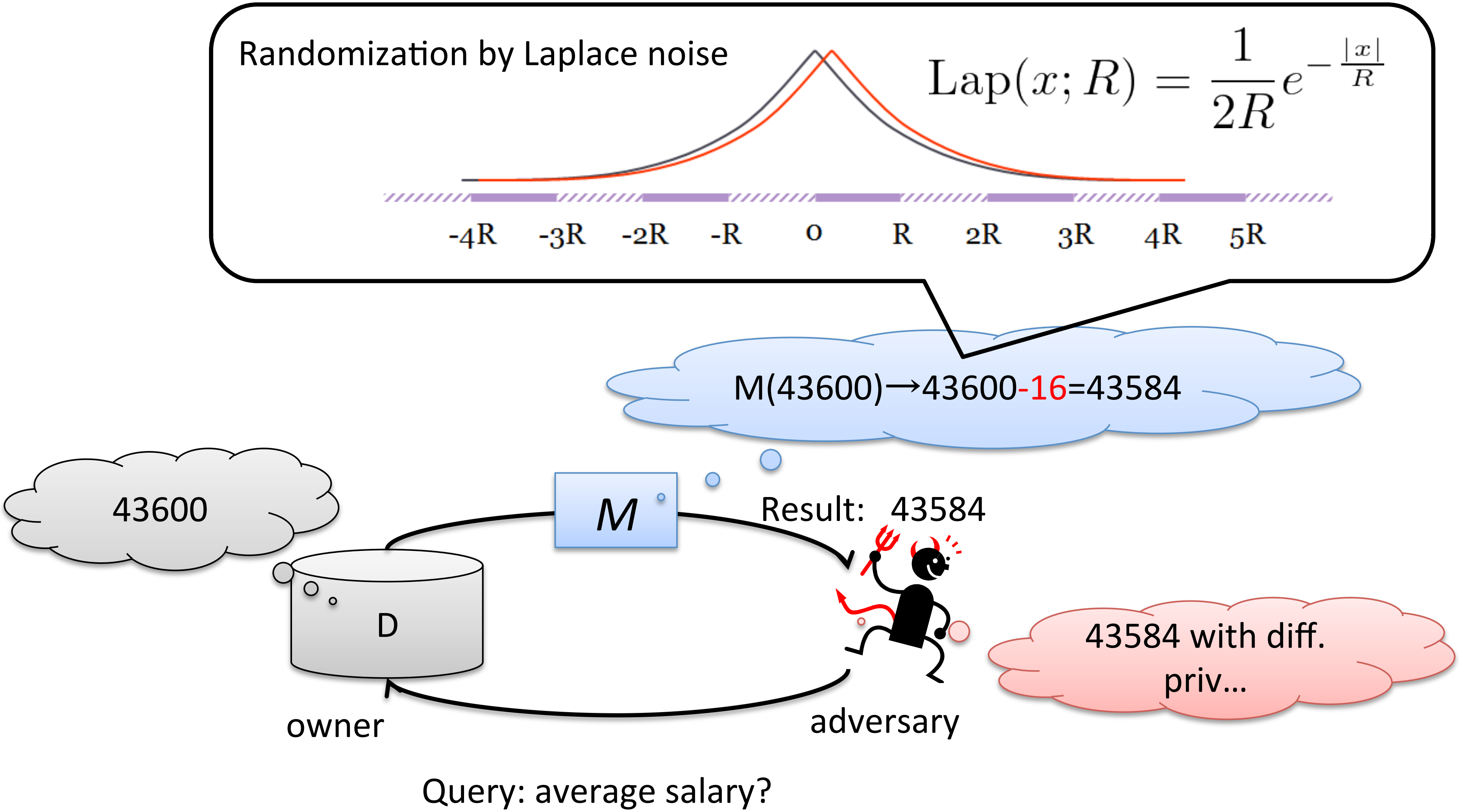
- **How can be define “not that different”?**

$$\frac{\Pr[f(\text{DB-You}) \in S]}{\Pr[f(\text{DB+You}) \in S]} \leq e^{\epsilon}$$

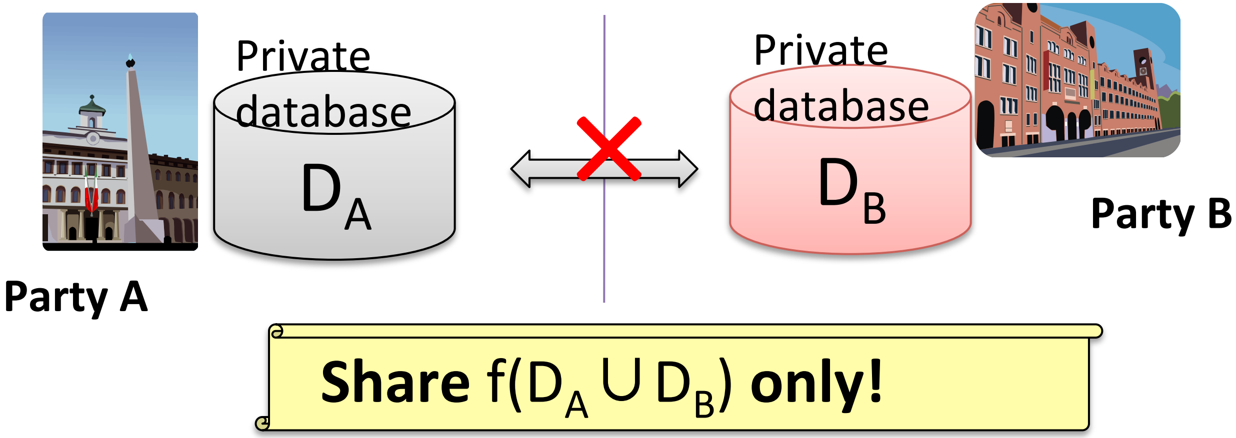


If the ratio of observing a response from D and D+A is close to 1, it is not privacy invasive.

Laplace Mechanism



Secure multiparty computation



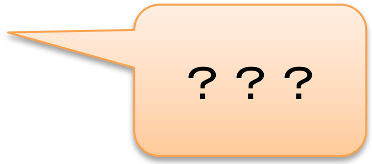
- Party A does not want to disclose D_a to B
- Party B does not want to disclose D_B to A, either
- But wish to learn compute $f(D_A \cup D_B)$
- Cryptographically secure computation

Computation over encrypted values

Holomorphic encryption

Bob 

Alice 



m_1, m_2 :secret numbers

c_1=Enc(m_1): encryption of m_1 → c_1

c_2=Enc(m_2): encryption of m_2 → c_2

$$\mathrm{Enc}(m_1)\cdot\mathrm{Enc}(m_2)=\mathrm{Enc}(m_1+m_2)$$

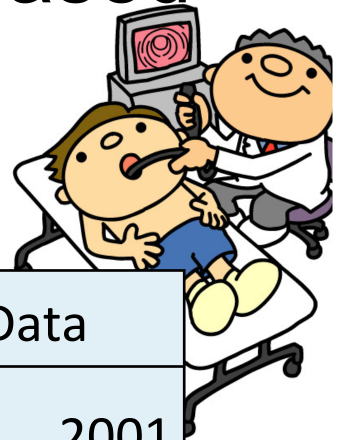
$$\mathrm{Enc}(m_1)^{m_2}=\mathrm{Enc}(m_1 m_2)$$

Privacy-preserving Epidemiology

- Estimate Relative Risk of cancer based on two medical data sets without revealing them

Name	Age	ROI
H. Kikuchi	25	Stomach
J. Sakuma	28	Stomach
J. Vaidya	35	lung

Hospital A
Patients of cancer



Name	Age	Data
M. Gofman	35	2001
H. Kikuchi	25	2001
C. Clifton	45	2002

Hospital B
Infection of H. pylori

Epidemiology study

- Contingency Table

H. Pylori	Cancer	None	Total
Yes	a	b	a/(a+b)
No	c	d	c/(c+d)

- Relative Risk (RR)

$$RR = \frac{\Pr(\text{cancer} \mid H.pylori)}{\Pr(\text{cancer} \mid \text{unexposed})} = \frac{a}{a+b} / \frac{c}{c+d} \approx \frac{ad}{bc}$$

Secure scalar product protocol

A (H.pylori)

X = (**Alice**,Bob,Carol,David)
= (1, 0, 1, 1)

B (cancer)

Y = (Alice,Bob,**Carol**,David)
= (1, 1, 1, 0)

Encryption X
E(1)

E(1), E(0), E(1), E(1) →

$c = E(1)^1 E(0)^1 E(1)^1 E(1)^0$
$= E(1*1)\ E(0*1)\ E(1*1)\ E(1*0)$
$= E(1 + 0 + 1 + 0)$
$= E(2)$

← c

Decryption
D(c) = D(E(2)) = 2
= | X ∩ Y |
number of common elements

H. Pylori	Cancer	None
Yes	2	
No		

Experimental Result

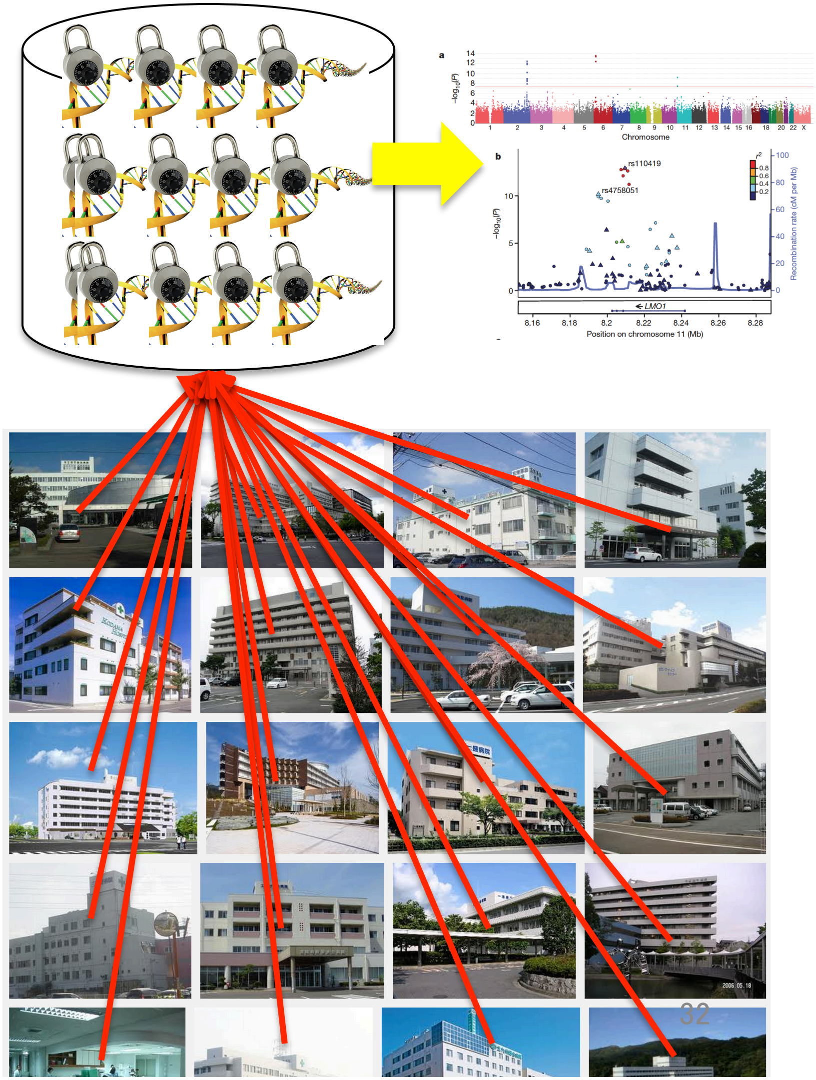
H. Pylori	Cancer	None	Total
Yes	80	2,549	2,629
No	346	106,988	107,334
total	426	109,537	109,963

Relative risk: RR = 9.70 (H. Pylori has 9.7times higher risk)

Test statistics χ = 17.8 (statistically significant with 98% conf.)

POINT: We have this without join of tables

Use private data as it is open and contribute to new scientific findings!

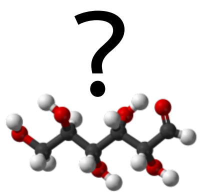


A use case: Secure pre-purchase inspection service

- The user wants to check whether or not the DB includes a sufficient number of similar compounds, w/o sending his/her private query.
- The server does not allow downloading of DB before purchase.

Dilemma!

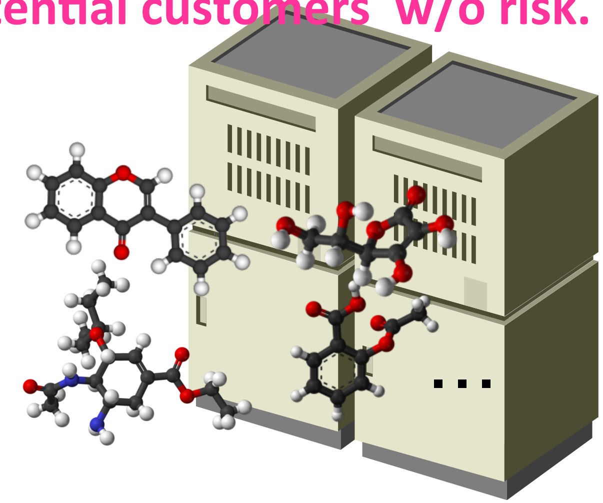
Benefit of using PPDM:
The client avoids “bad investment” w/o risk.
DB obtains more potential customers w/o risk.



A query

How many similar compounds?

Num. of similar compounds



Evaluating Similarity by AHE

$$S_{Jacc} = \frac{|p \cap q|}{|p \cup q|} > \frac{\theta_n}{\theta_d} \quad \leftrightarrow \mathrm{T}_{p,q} := (\theta_n + \theta_d)\,|p \cap q| + \theta_n(-|p|) + \theta_n(-|q|) \geq 0$$

$T_{p,q} > 0$, iff $S_{jacc} > \theta_n/\theta_d$

$T_{p,q}$ is computed by additive operations only.

$p = (p_1,\ldots,p_n)$

$Enc(q_1),\cdots,Enc(q_n)$

$Enc(-|q|)$

$\longleftarrow$

$Enc(T)$

$\longrightarrow$

$q = (q_1,\ldots,q_n)$

The user checks
T>0 or not.

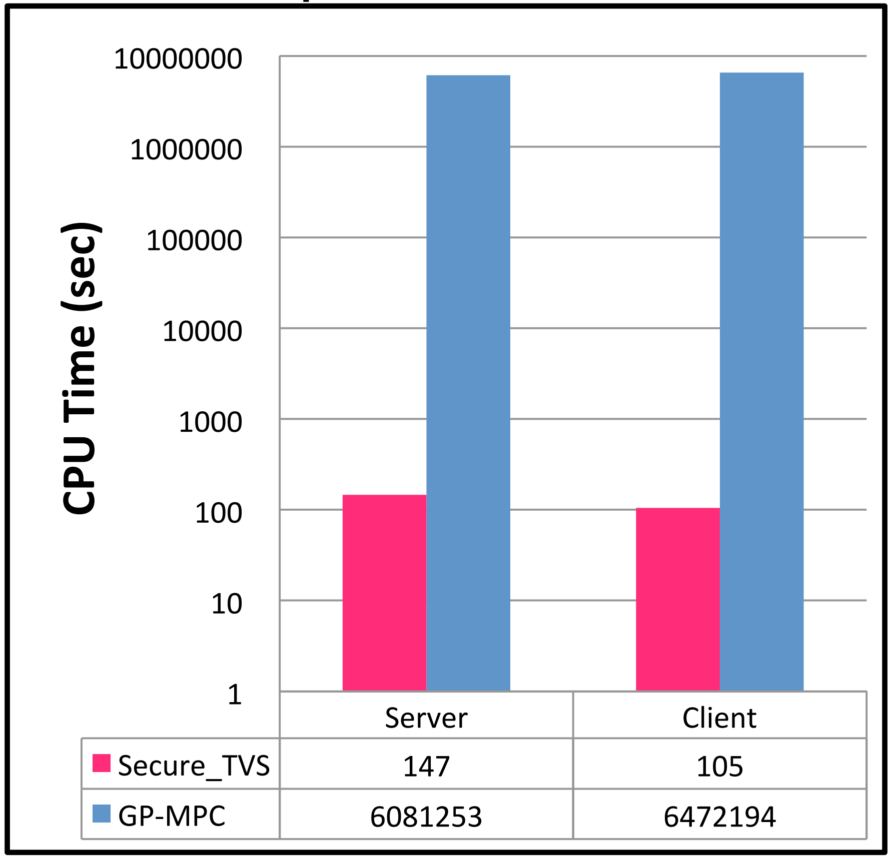
$$Enc(T) = Enc(\lambda_1|p \cap q| + \lambda_2|-p| + \lambda_3|-q|) =$$

$$\underbrace{Enc(|p \cap q|) \oplus \cdots}_{\lambda_1} \oplus \underbrace{Enc(|-p|) \oplus \cdots}_{\lambda_2} \oplus \underbrace{Enc(|-q|) \oplus \cdots}_{\lambda_3}$$

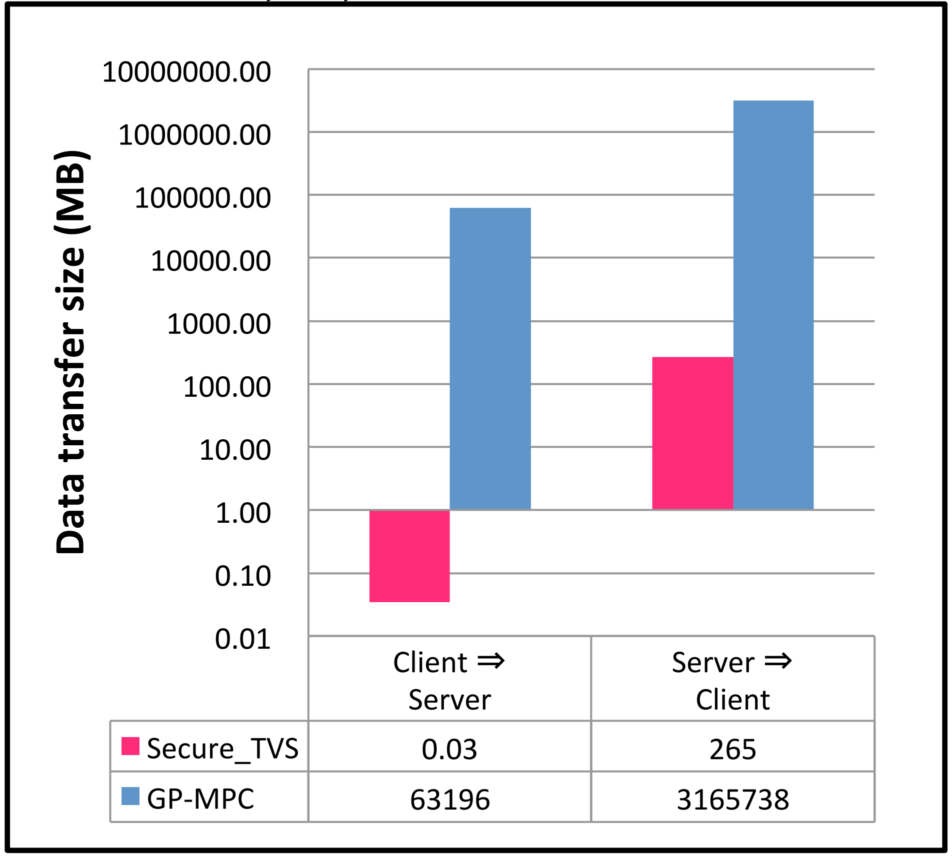
Shimizu et al., bioRxiv (2015)

Results

- A dataset: Latest ChEMBL (1,292,344 compounds)
- CPU: Xeon 2.9GHz
- Compared to GP-MPC (The Fairplay system)

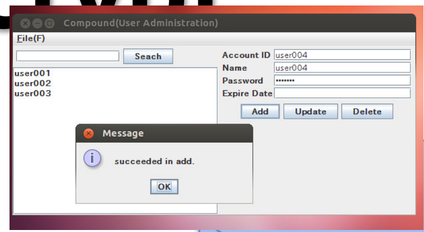


50,000 times faster computation

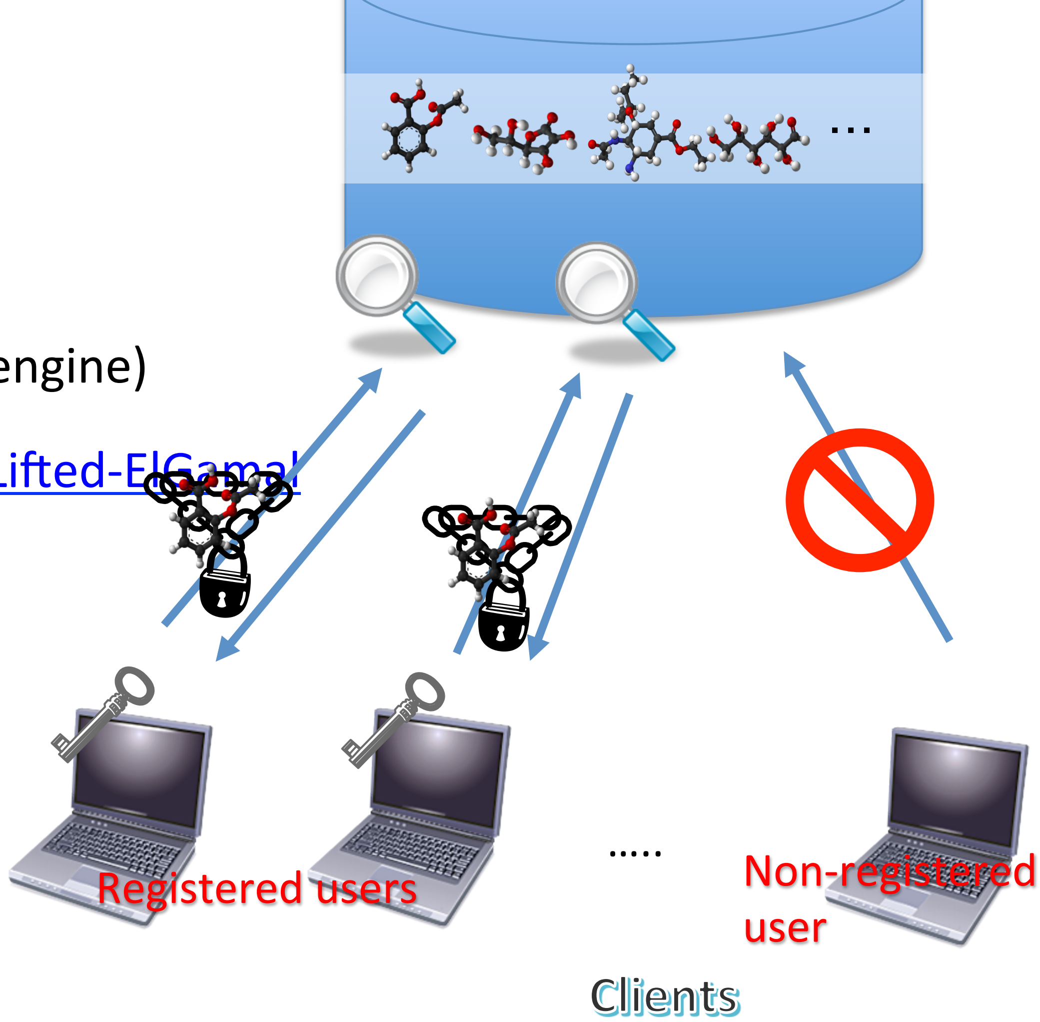


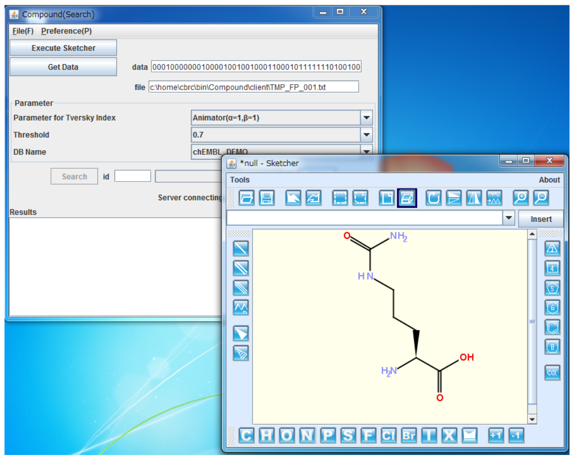
12,000 times smaller data transfer

ChemCrypt



- Server & Client system
  - PP-Similarity Search
  - User Authentication System
  - Parallel Processing (Server)
  - User Friendly GUI
    - Compound Sketcher (Client)
- Implementation
  - Java (GUI, DB, etc), C++ (core engine) using AistCrypt https://github.com/aistcrypt/Lifted-ElGamal





Personalized medication with preserving genomic privacy

Jun Sakuma
U. Tsukuba CS dept./JST CREST

What we can see from personal genome

Strongly associated with genetic information

- Genetic disorder, type-2 diabetes (2型糖尿病)
- Origin of ancestors
- Sensitivity to alcohol
- Risks at common diseases (heart attack, hyper tension, fatness, etc.)
- Athletic ability, IQ, character

Weakly associated with genetic information

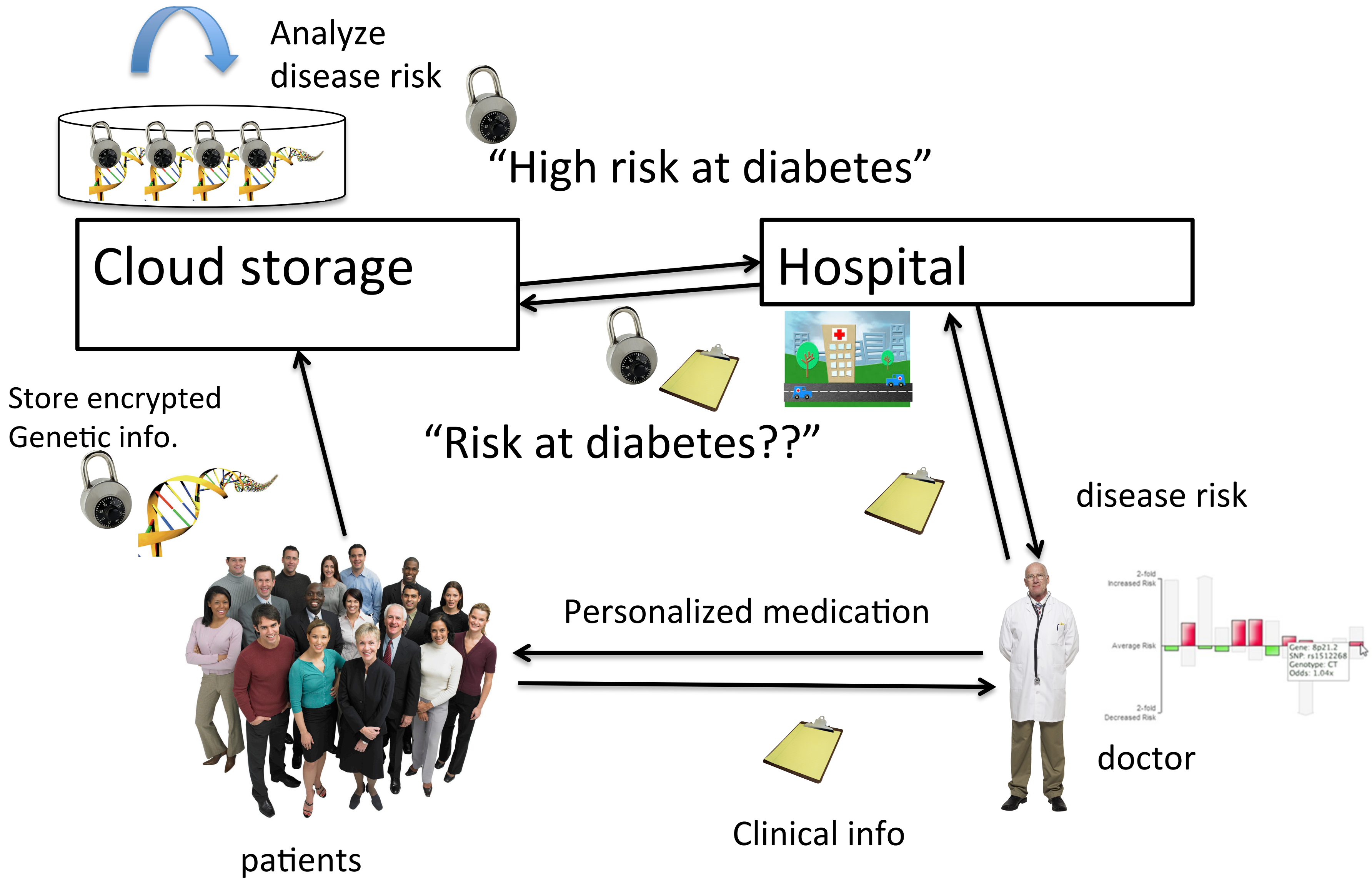
Genome-based personalized medicine

- Preventive medication
  - Estimation of risks at common diseases
    - Improvement of meals and daily habits
  - Angelina Jolly's case
    - (According to a genetic test) "My doctors estimated that I had an 87 percent risk of breast cancer and a 50 percent risk of ovarian cancer, although the risk is different in the case of each woman."
    - Angelina Jolly replaced her breast with implant
- Drug administration
  - Anti-cancer drug selection
  - Adjustment of dosage amount
- Change the quality of medicine drastically

Genomic privacy

- Personal genome can be applied to various types of medical tests
  - Drug administration
  - Disease risks at common diseases
- Risks of personal genome utilization
  - If leaked, it may reveal genetic disorder, potential disease risks, abilities, IQ, and so on.
  - can cause serious discrimination: health insurance, hiring, marriage...
- We need a framework to analyze personal genome securely and privately

Privacy-preserving personalized medication



Risk Evaluation from personal genome

$$\text{risk} = \boldsymbol{w}^T \boldsymbol{x} = \overbrace{w_1^G x_1^G + w_2^G x_2^G}^{\text{Risk from genetic factors}} + \ldots + \overbrace{w_1^C x_1^C + w_2^C x_2^C}^{\text{Risk from clinical factors}} + \ldots$$

Presence of gene1

effect of gene1 on the risk
of developing DT

A clinical factor (e.g. age)

effect of the factor on the risk
of developing DT

Privacy-preserving personalized medication

- Disease risk evaluation

$$\boldsymbol{w}^T\boldsymbol{x} = \sum_i w_i^G x_i^G + \sum_i w_i^C x_i^C$$

- Disease risk evaluation with enc. values

$$\begin{aligned} enc(\boldsymbol{w}^T\boldsymbol{x}) &= enc(\sum_i w_i^G x_i^G + \sum_i w_i^C x_i^C) \\ &= \prod_i enc(w_i^G x_i^G) \cdot \prod_i enc(w_i^C x_i^C) \\ &= \prod_i enc(x_i^G)^{w_i^G} \cdot \prod_i enc(x_i^C)^{w_i^C} \end{aligned}$$

My results were…

- Used 20 genetic factors, 4 clinical factors
- risk at diabetes (糖尿病)
  - 10.9 % from the bottom
  - Very low
- Risk at hypertension (高血圧)
  - 45.8% from the bottom
  - Quite average
- Note that I could have the results without sharing my genetic/clinical information with anybody