Preserving Privacies in Biomedical Data with "More Efficient" Differentially Private Algorithms

Tetsuo Shibuya

Division of Medical Data Informatics,
Human Genome Center, Institute of Medical Science,
The University of Tokyo

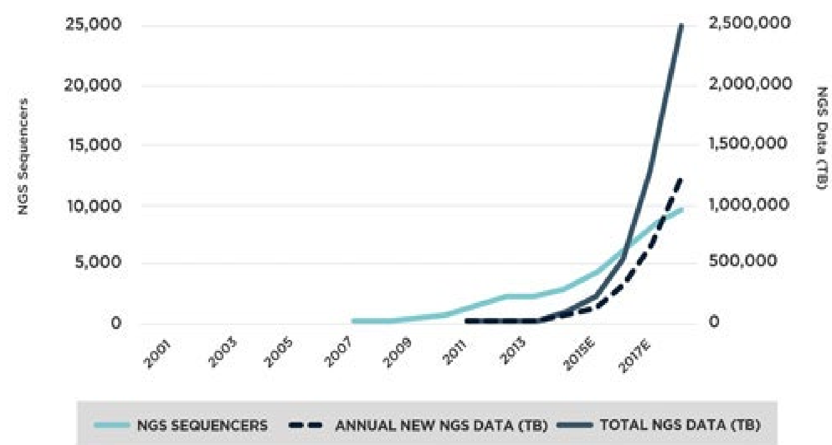
Today's Talk

- Differential privacy algorithms and its applications
  - Techniques
    - Sensitivity analysis
    - Bias reduction
    - Multiple attributes
    - *k*-anonymized differential privacy
  - Applications
    - Genome-wide association study
    - Graph databases

Next Generation Sequencers (NGS)

T. Shibuya

- One of the greatest innovation in genome science
  - Fast:
    - 8Tbp / 1 day (Illumina NovaSeq X)
      - ～60 individuals per day
  - Cheap:
    - 200–300 dollars per individual



[Davis-Dusenbery, 2017]

Genome Data Explosion

cf. Costed 3 billion dollars and 13 years in the Human Genome Project (～2003)

Illumina
NovaSeq X

https://jp.illumina.com/systems/sequencing-platforms/novaseq-x-plus.html

Oxford Nanopore
MinION

https://nanoporetech.com/sites/default/files/s3/minion-usb.png

PacBio
Sequel II

https://www.pacb.com/products-and-services/sequel-system/latest-system-release/

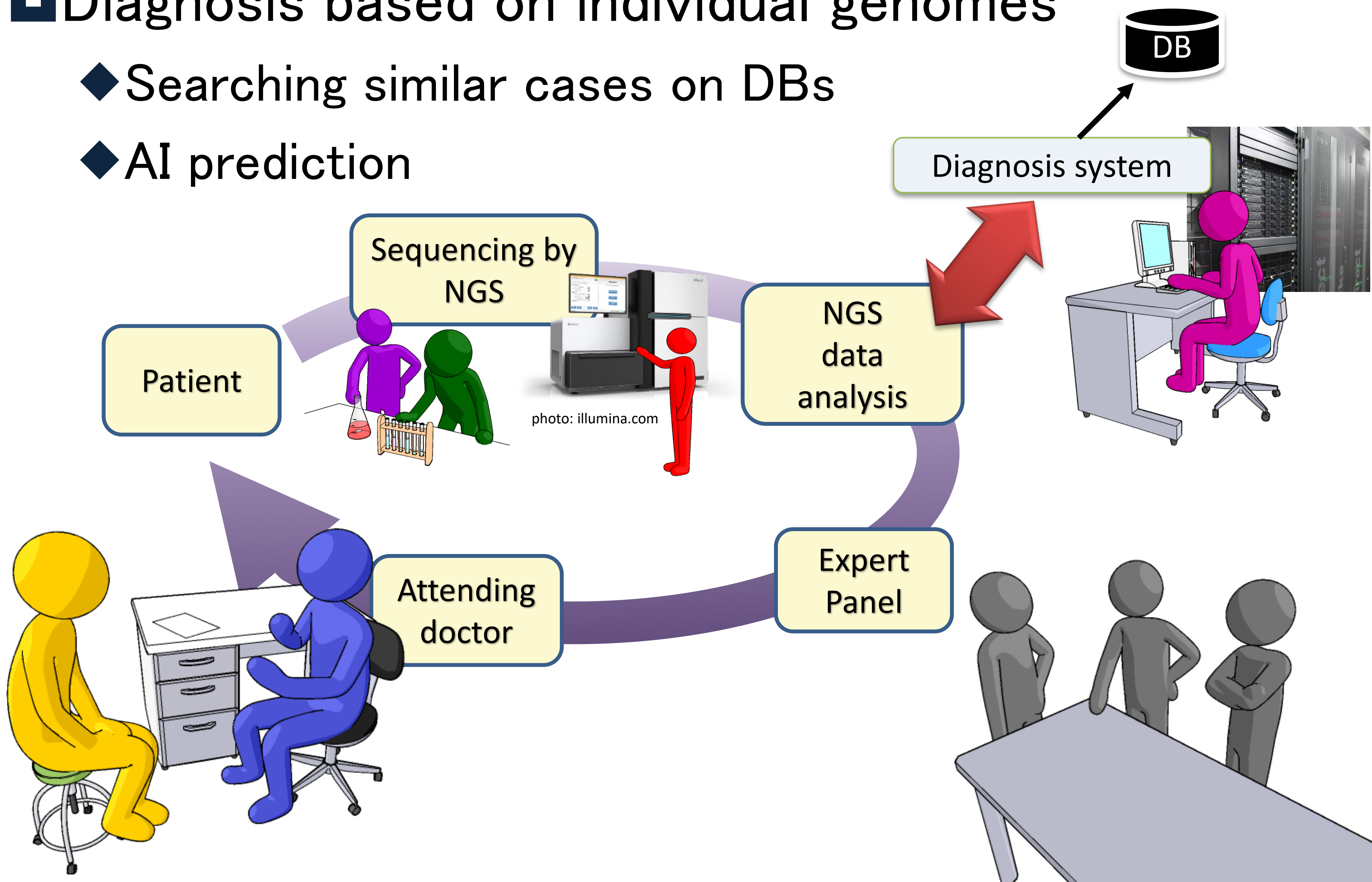
MGI
DNBSEQ-T20x2

https://jp.mgi-tech.com/products/instruments_info/22/

Clinical Sequencing for Precision (Personalized) Medicine

T. Shibuya

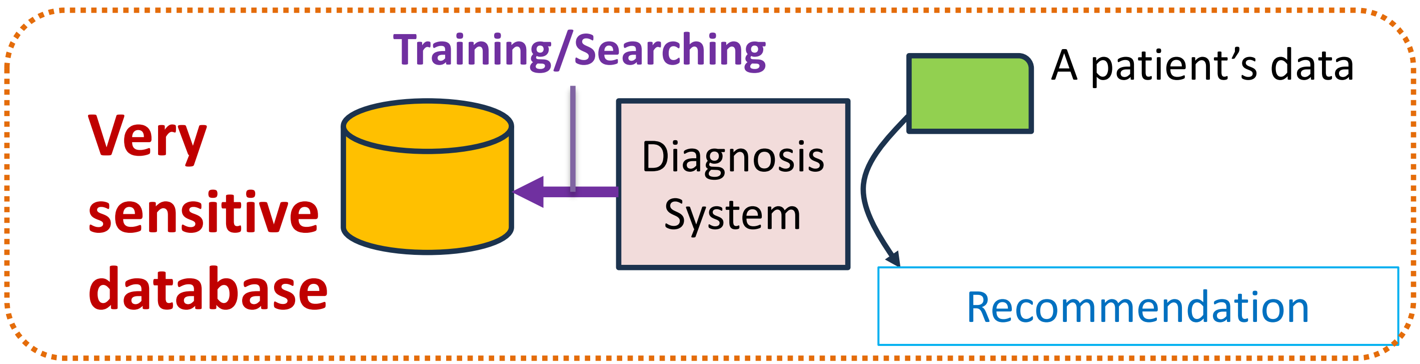
- Diagnosis based on individual genomes
  - Searching similar cases on DBs
  - AI prediction



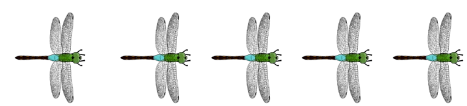
A Concern on Precision Medicine

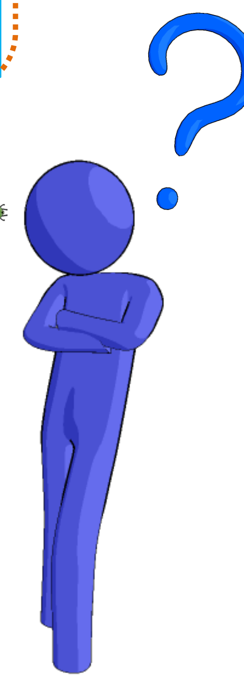
T. Shibuya

- **It utilizes highly sensitive data**
  - Including genomes of other people



Question



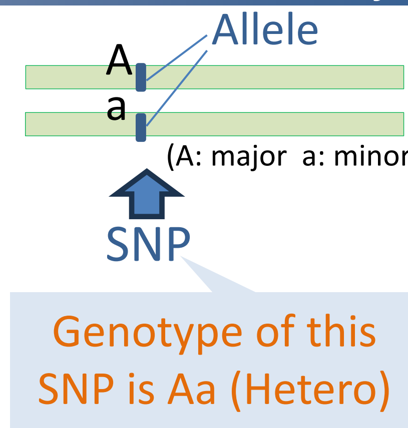


Can we assure that the recommendation does not contain any sensitive private information?

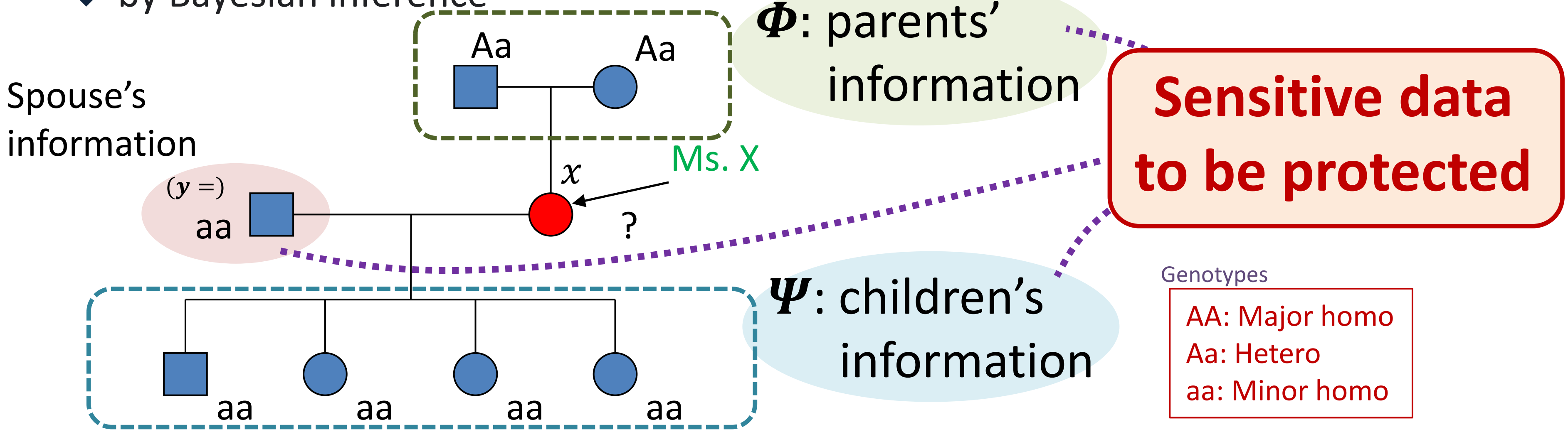
Example: Genotype Analysis using Pedigrees

T. Shibuya

- Probabilities of possible genotypes x of Ms. X
  - From only the parent information:
    - $P(x = \mathrm{AA}|\Phi) = 1/4, P(x = \mathrm{Aa}|\Phi) = 1/2, P(x = \mathrm{aa}|\Phi) = 1/4$
  - Probabilities of children's genotypes under 3 possible cases
    - $P(\Psi|x = \mathrm{AA}, y = \mathrm{aa}) = 0$
    - $P(\Psi|x = \mathrm{Aa}, y = \mathrm{aa}) = 1/16$
    - $P(\Psi|x = \mathrm{aa}, y = \mathrm{aa}) = 1$



- Hence $P(x = \mathrm{AA}) = 0, P(x = \mathrm{Aa}) = 1/9, P(x = \mathrm{aa}) = 8/9$
  - by Bayesian inference



How secure is it to disclose the probabilities (i.e., 0, 1/9, 8/9) to Ms. X?

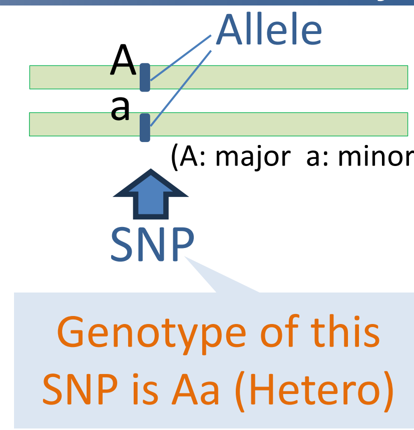
Technical Terms

- DNA
  - Chain polymer molecule composed of 4 types of nucleic acids: A/T/C/G
- Chromosomes
  - DNA molecules in a cell
  - We have 23 pairs of chromosomes (1–22 and X/Y)
- SNPs (Single nucleotide polymorphisms)
  - Specific positions with single nucleotide variations
- Alleles
  - Type of the nucleic acid at the SNP
- Major/Minor Alleles
  - The most common type of a SNP is called the major allele
  - Other types are called minor alleles
- Genotypes
  - Pair of alleles at the SNP
  - Called homozygous (or homo) if both alleles are the same
    - major homo/minor homo
  - Called heterozygous (or hetero) otherwise

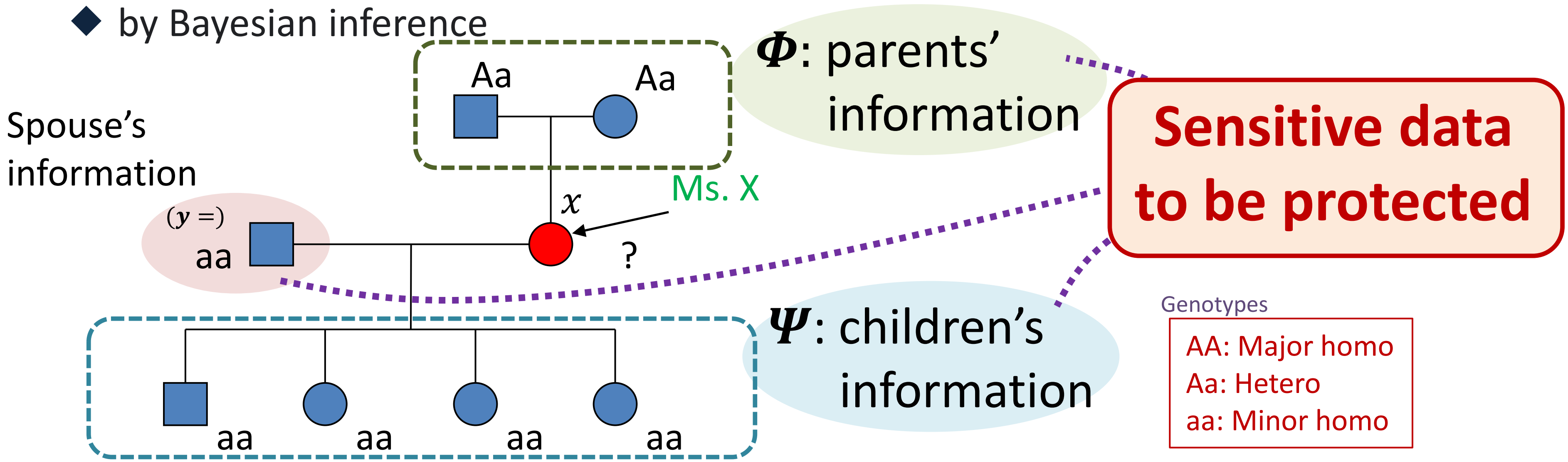
Example: Genotype Analysis using Pedigrees

T. Shibuya

- Probabilities of possible genotypes x of Ms. X
  - From only the parent information:
    - $\mathrm{P}(x = \mathrm{AA}|\Phi) = 1/4, \mathrm{P}(x = \mathrm{Aa}|\Phi) = 1/2, \mathrm{P}(x = \mathrm{aa}|\Phi) = 1/4$
  - Probabilities of children's genotypes under 3 possible cases
    - $\mathrm{P}(\Psi|x = \mathrm{AA}, y = \mathrm{aa}) = 0$
    - $\mathrm{P}(\Psi|x = \mathrm{Aa}, y = \mathrm{aa}) = 1/16$
    - $\mathrm{P}(\Psi|x = \mathrm{aa}, y = \mathrm{aa}) = 1$



- Hence $\mathrm{P}(x = \mathrm{AA}) = 0, \mathrm{P}(x = \mathrm{Aa}) = 1/9, \mathrm{P}(x = \mathrm{aa}) = 8/9$
  - by Bayesian inference



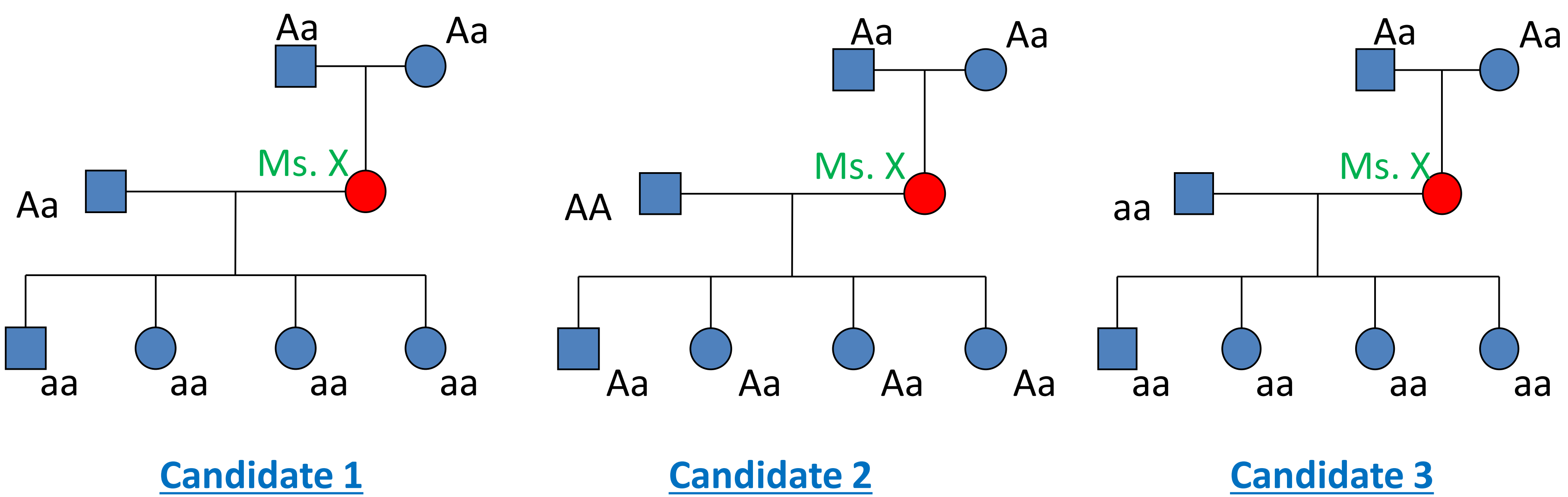
Is it OK to disclose the probabilities (i.e., 0, 1/9, 8/9) to Ms. X?

It leaks much information!!

- Given:
  - Probabilities of Ms. X's possible genotypes are:

$$\mathrm{P}(x = \mathrm{AA}) = 0, \mathrm{P}(x = \mathrm{Aa}) = 1/9, \mathrm{P}(x = \mathrm{aa}) = 8/9.$$

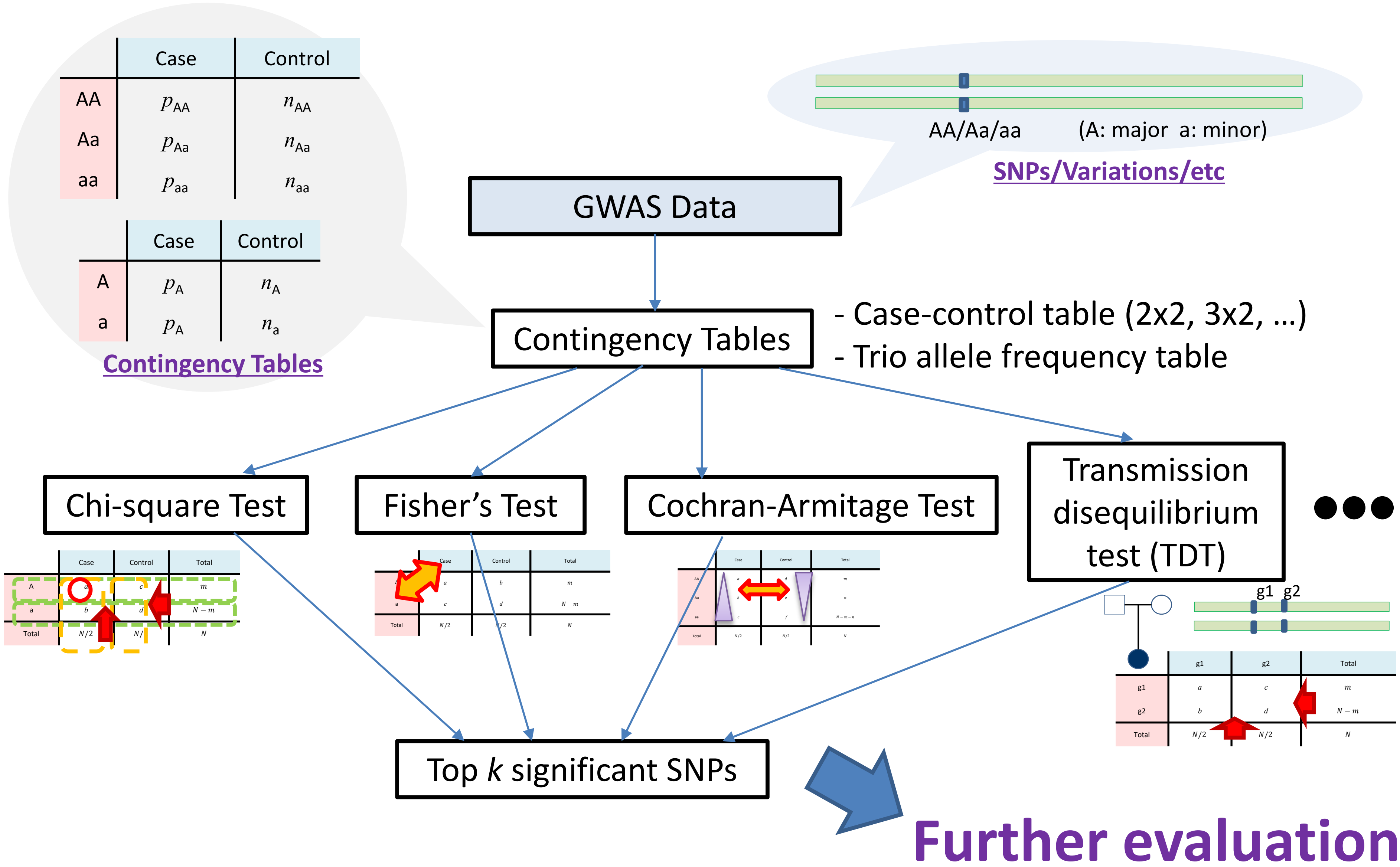
- Only 3 possible cases (as below) exist, which means:
  - Her parents' genotypes are revealed
  - All the other genotypes are also be revealed, if she know her husband's genotype



GWAS (Genome-Wide Association Study)

T. Shibuya

■ Statistical analyses for finding important genes/SNPs/etc.



Statistics used in GWAS

T. Shibuya

No problem to publish these statistics?

- χ^2 test
  - $\chi^2 = \dfrac{N(2a-m)^2}{m(N-m)}$
- Fisher's independence test
  - $p = \dfrac{\binom{m}{a}\binom{N-m}{c}}{\binom{N}{m}}$
- Cochran–Armitage's trend test
  - $\chi^2 = \dfrac{N(2m+n-2(2a+b))^2}{N(4m+n)-(2m+n)^2}$
- Top k significant genes
  - Output genes with the k largest test values

Predictable?

	Case	Control	Total
A	a	c	m
a	b	d	$N-m$
Total	$N/2$	$N/2$	N

Independent?

	Case	Control	Total
A	a	b	m
a	c	d	$N-m$
Total	$N/2$	$N/2$	N

	Case	Control	Total
AA	a	d	m
Aa	b	e	n
aa	c	f	$N-m-n$
Total	$N/2$	$N/2$	N

=?

Contingency Tables

Differential Privacy [Dwork 2006]

Gödel Prize 2017

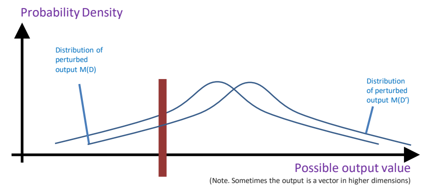
■ **Noise addition strategy** for preserving privacy

◆ **Differential privacy** is satisfied if:

One entry difference
(insertion/deletion/update)

Database D

Database D’



Noise-added analysis

based on DP

Result M(D)

Any kind of results
(Statistics, prediction result, etc)

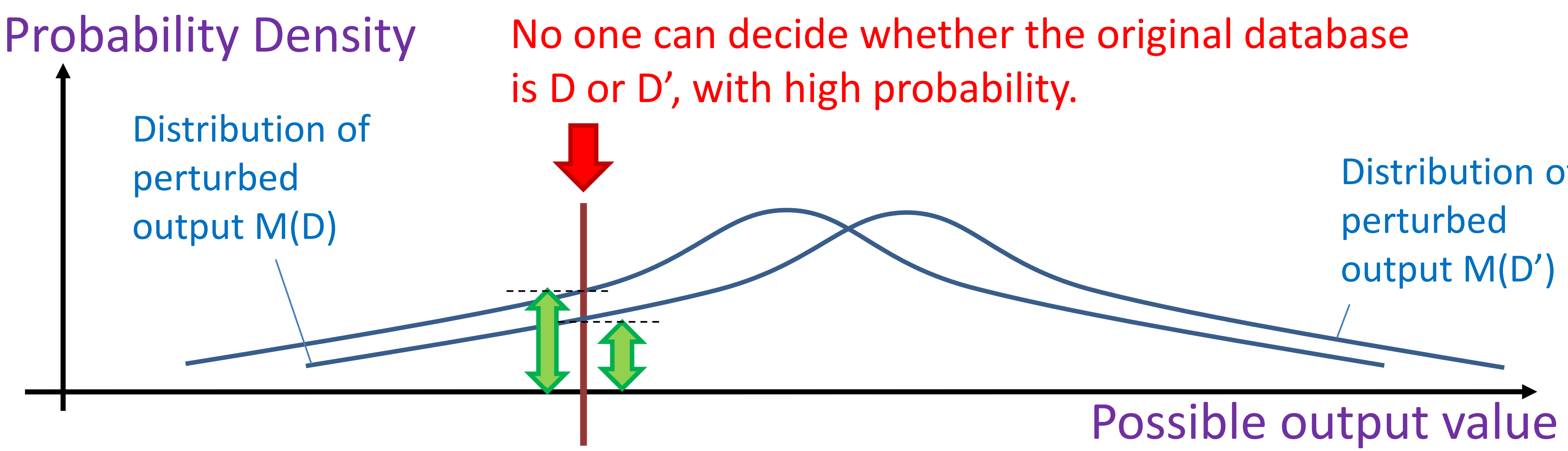
Result M(D’)

M(D) and M(D’) is
probabilistically indistinguishable.

ε-Differential Privacy [Dwork 2006]

T. Shibuya

- Noise mechanism M is said to be ε –differentially private *iff*
  - for any two databases D and D' s.t., $|D-D'|=1$
    - *i.e.*, one entry difference
  - for any output set S
    - $\Pr[M(D) \in S] \le e^{\varepsilon} \cdot \Pr[M(D') \in S]$
      - ε: **Privacy budget**

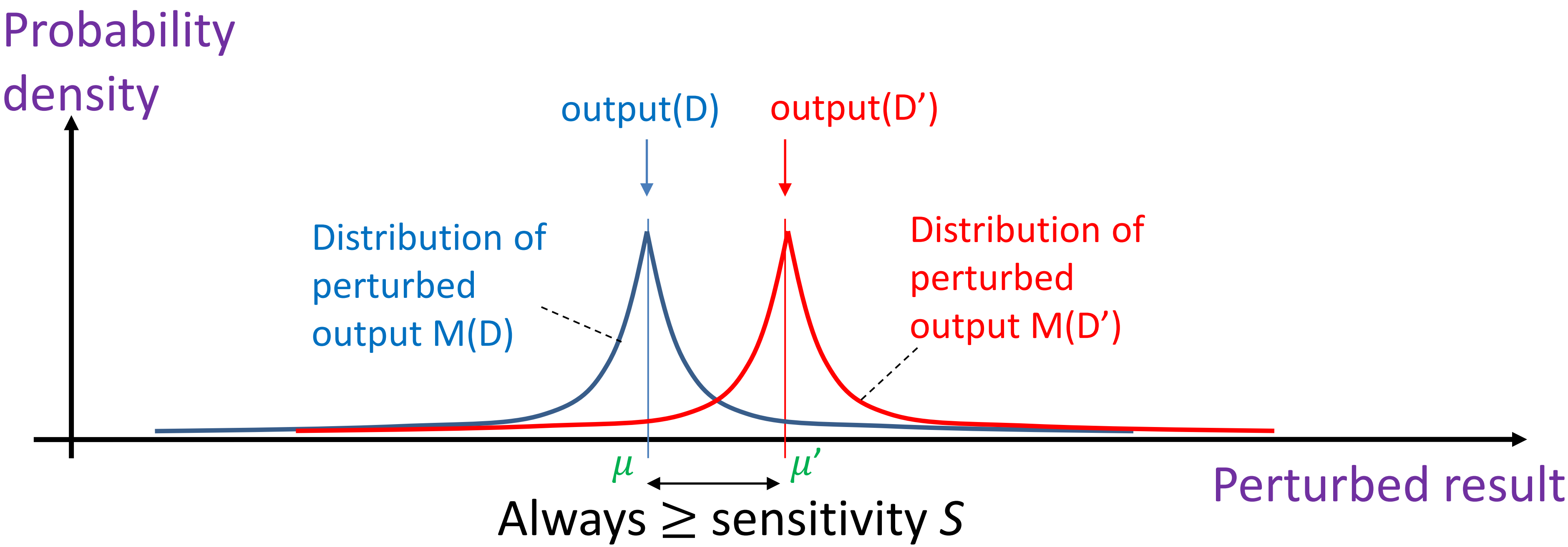


(Note. Sometimes the output is a vector in higher dimensions)

Example: Laplace Mechanism [McSherry, Talwar, 2007]

T. Shibuya

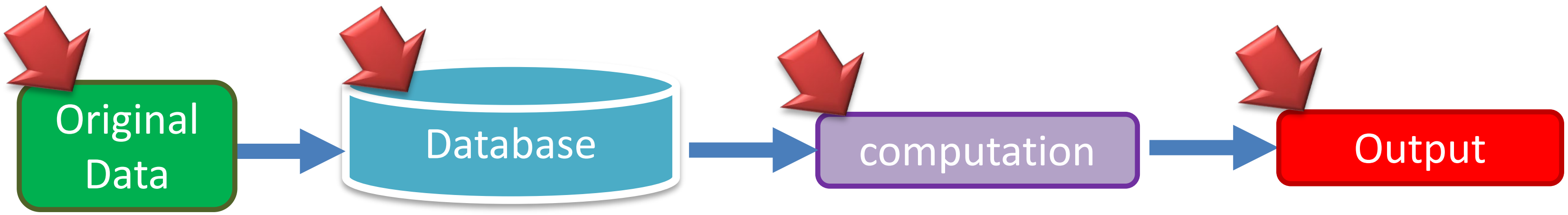
- Add noises following the Laplace distribution:
  - $\frac{\varepsilon}{2S} e^{\left(-\frac{|x-\mu|}{S}\varepsilon\right)}$
    - μ : actual output ε: parameter for ε−differential privacy
    - S: Sensitivity (Minimum difference between output(D) and output(D')) ($|D$-$D'|$=1)
- Then
  - $\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S]$ for any D and D'



Two Important Properties of the Differential Privacy

Flexible applications

- Noise can be added at any stage
  - Local data before uploading / database / algorithm inside / output results / trained parameters / etc.



- **More noise**
- **For general use**
- **Easier to Design**

Trade-offs

- **(Possibly) less noise**
- **For specific use**
- **Difficult to design (in general)**

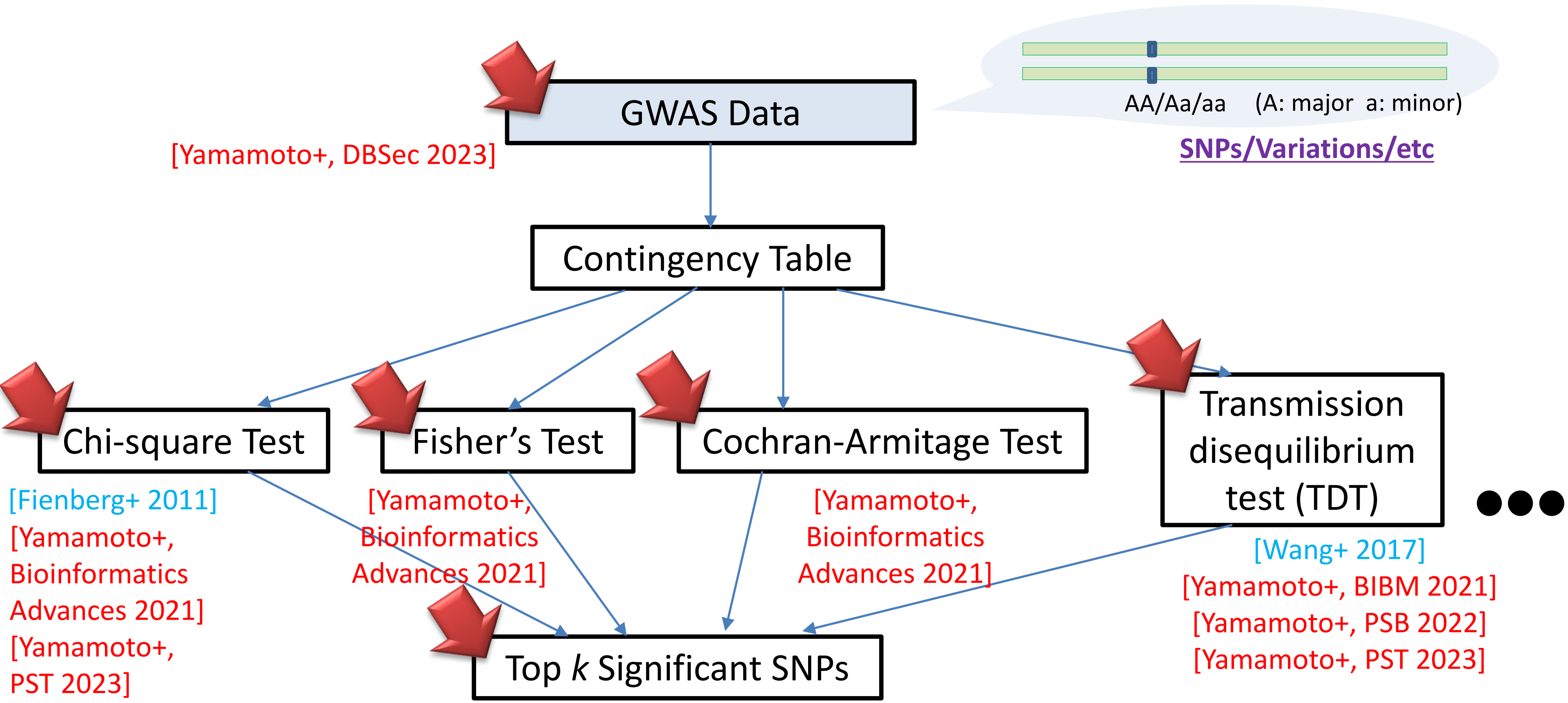
Robustness against attacks

- Any postprocessing on already ε−differentially private data is kept to be ε−differentially private
  - i.e., Theoretically no one can break ε−differential privacy!

Our Recent Research: DP mechanisms for GWAS

T. Shibuya

Differentially Private Mechanism Design for GWAS

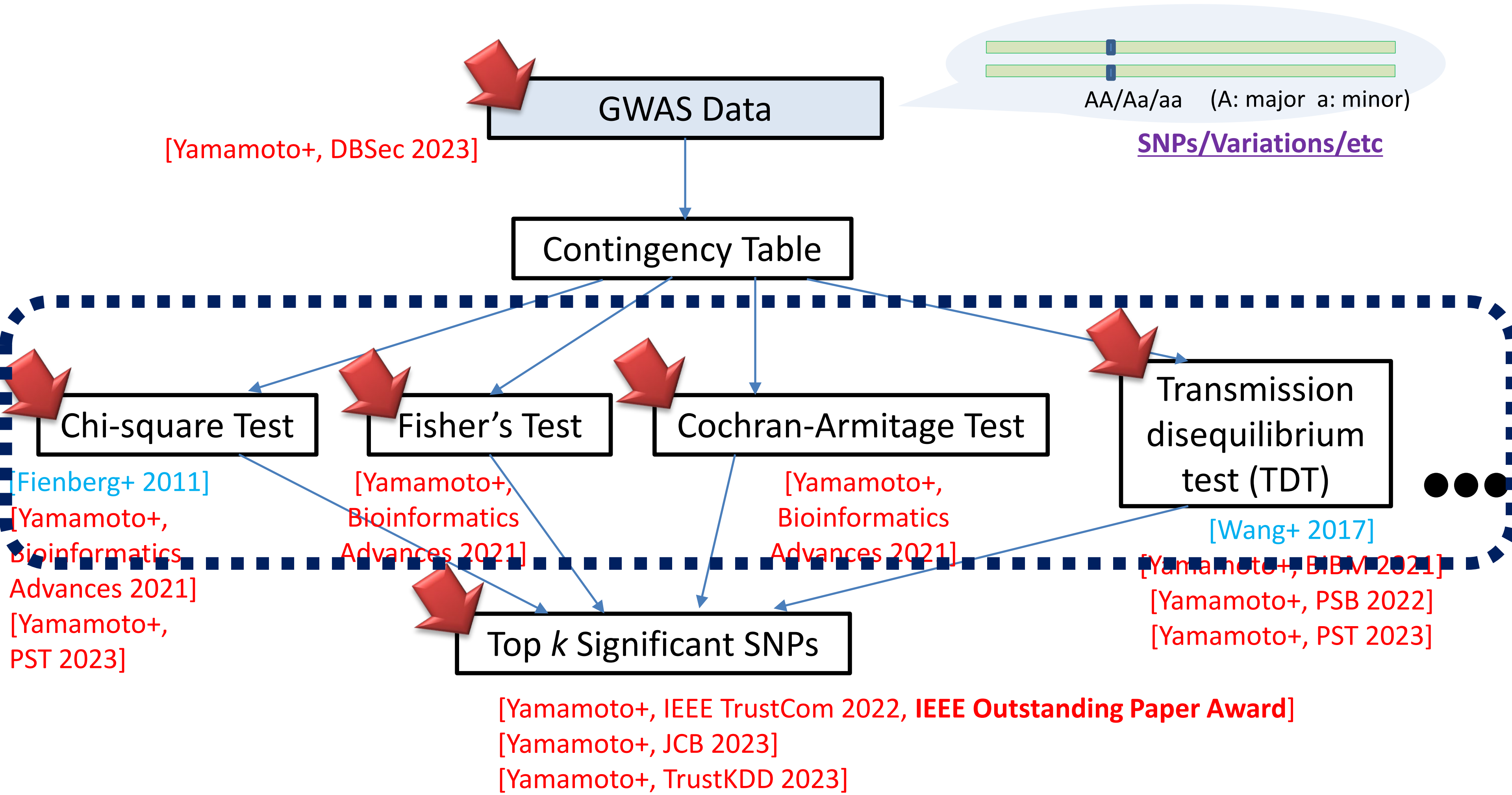


[Yamamoto+, IEEE TrustCom 2022, **IEEE Outstanding Paper Award**]
[Yamamoto+, JCB 2023]
[Yamamoto+, TrustKDD 2023]

Our Recent Research: DP mechanisms for GWAS

T. Shibuya

■ Differentially Private Mechanism Design for GWAS



Sensitivity Analyses for Laplace Mechanism for GWAS Tests

T. Shibuya

- Sensitivity of the χ^2 test
  - $S = 4N/(N+4)$ [Fienberg+ 2011]
  - Sensitivity of $\log_{10}$(P–value)
    - ▶ 2.33 (i.e., constant)
- Sensitivity of the Fisher's independence test
  - $S = \dfrac{N(7N-6)}{32(N-1)(N-3)}$
- Sensitivity of the Cochran–Armitage's trend test
  - $S = \dfrac{16N(N^2+6N+4)}{(N+18)(N^2+8N-4)}$

…

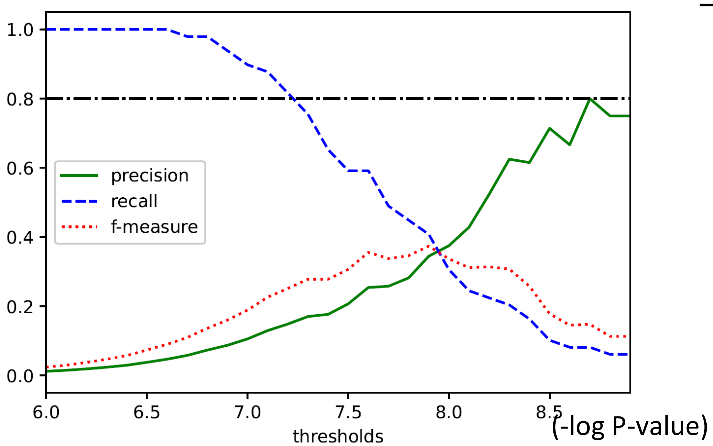
Our Result [Yamamoto+, Bioinfor. Adv. 2021]

	Case	Control	Total
A	a	c	m
a	b	d	$N-m$
Total	$N/2$	$N/2$	N

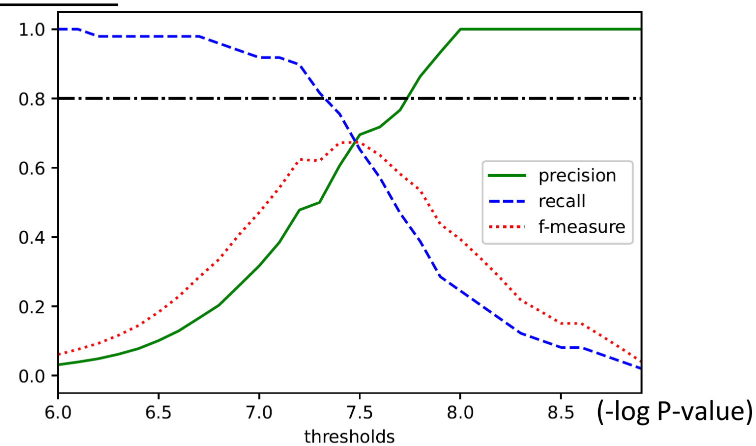
	Case	Control	Total
AA	a	d	m
Aa	b	e	n
aa	c	f	$N-m-n$
Total	$N/2$	$N/2$	N

Contingency Tables

$\varepsilon = 7.0$



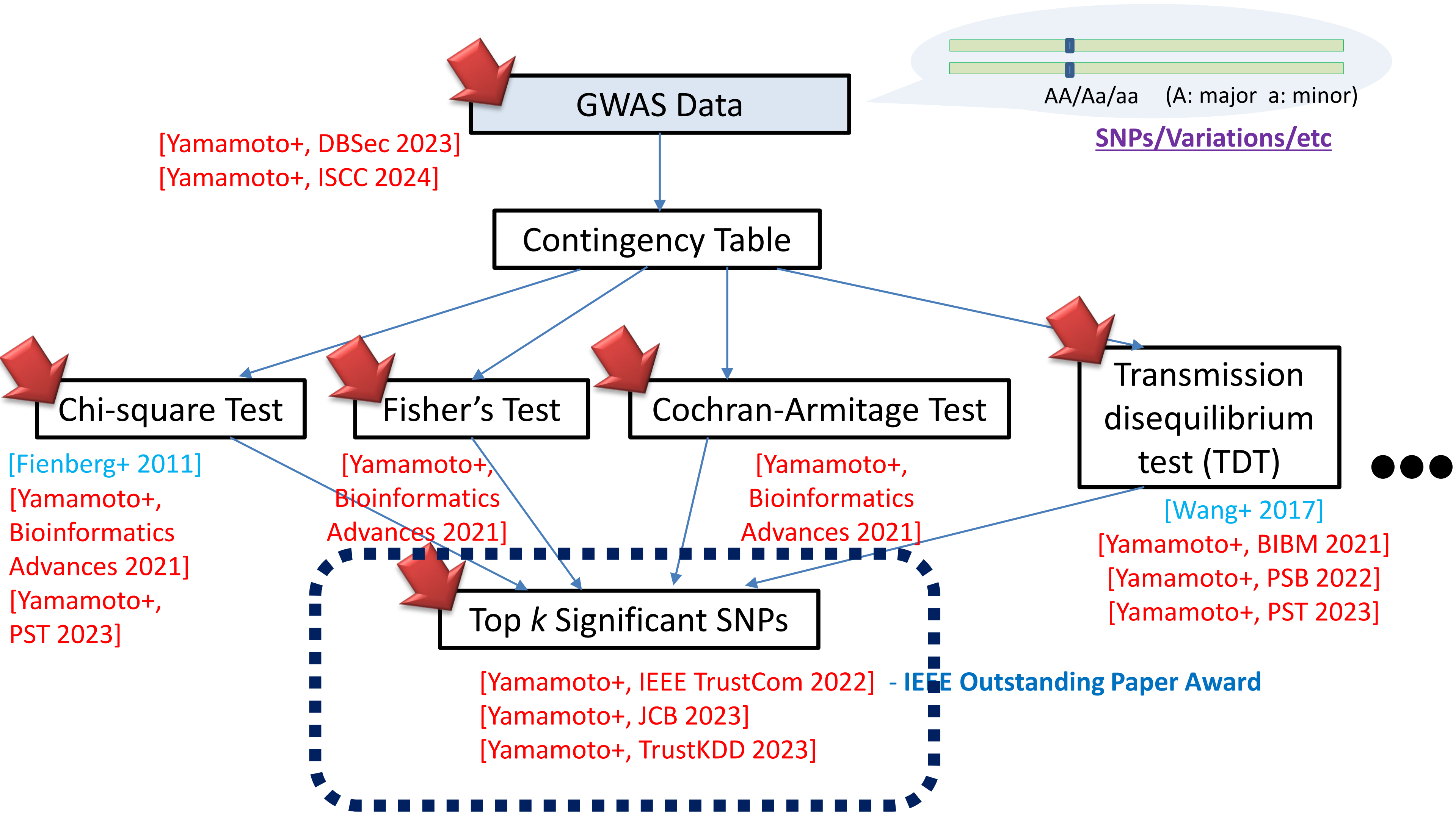
$\varepsilon = 10.0$



Trade-off between privacy and accuracy in Fisher's Test

Our Recent Research: DP mechanisms for GWAS

■ Differentially Private Mechanism Design for GWAS



The top k significant SNPs/genes/etc

- We can obtain the DP top k significant SNPs by adding DP noise to each SNP value, but it does not work well.
  - as we need to add $\sqrt{n}$ times larger noise – **TOO LARGE!**
    - than the case of publishing a single SNP result
    - n: Number of SNPs

$-\log P$ **values**	**SNPs**
103.55	X
87.64	Y
53.37	Z
49.55	W
47.32	V
42.20	U
…	…

SNPs sorted by P-values

output the 3-most significant genes

X, Y, Z

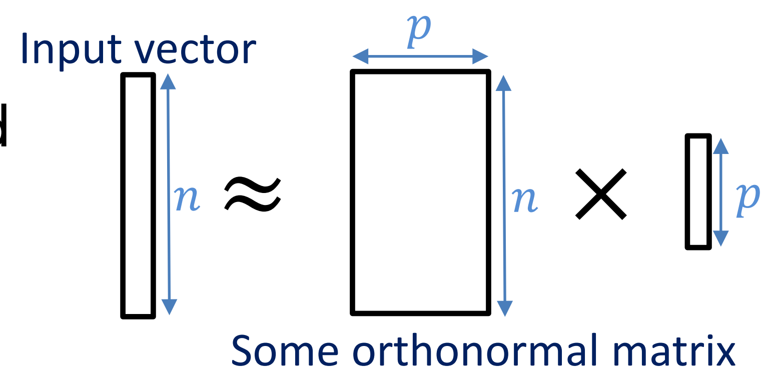
Observation

- We can reduce it to $O(\sqrt{k})$ in case we publish only k specific pre-determined SNPs data.
  - $k \ll n$
  - But we cannot know which to publish beforehand

$-\log P$ values	SNPs
103.55	X
87.64	Y
53.37	Z
49.55	W
47.32	V
42.20	U
…	…

Output P-values of 3 Specific SNP data

Compressive Mechanism [Li et al., 2011]

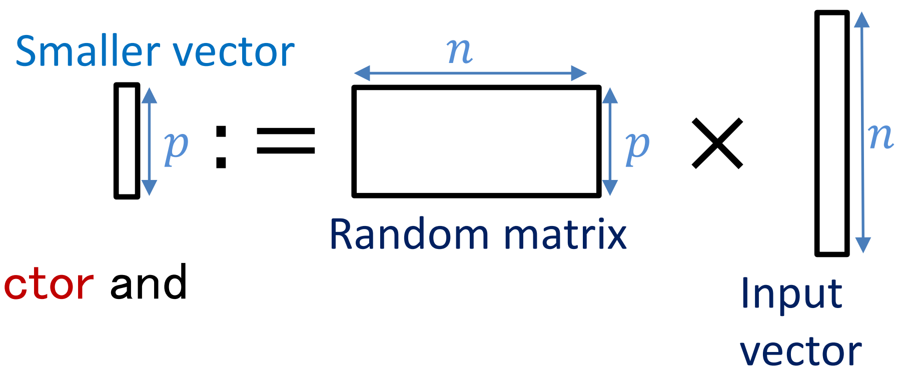


Assumption

- It can be approximately represented by a very small vector

Compressed sensing

- Find a representative small vector by random projection
  - Appropriate p must be given



Compressive Mechanism

- Add noise to the smaller projected vector and reconstruct the original vector!
- Pros
  - Noise size can be $O(\sqrt{p})$ instead of $O(\sqrt{n})$
- Cons
  - Very slow, and it cannot be applied to the entire SNPs data
  - Works well only for only sparse data
    - Could contain more errors if not

Our Mechanism for Publishing top-k SNPs

[RECOMB Genome Security Workshop, 2022 (JCB 2023)]

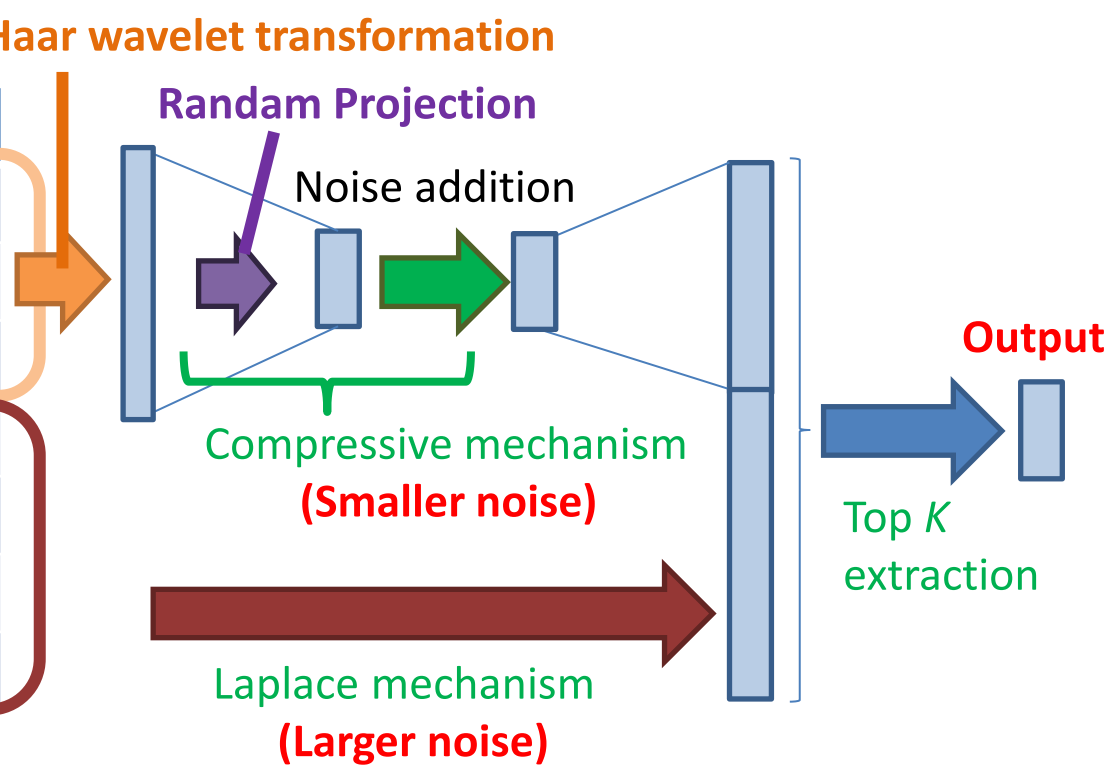
T. Shibuya

- Enhanced compressive mechanism
  - Add smaller noise to top-rank SNPs by compressive mechanism
    - after sparsification by Haar wavelet transformation
  - Add Laplace noise to other SNPs
  - Merge them and extract top k SNPs
    - 2x noise needed, but still better than just applying only Laplace mechanism
- The output is still ε-differentially private

k' ($k' \geq k$)

$-\log P$ values	SNPs
103.55	X
87.64	Y
53.37	Z
49.55	W
47.32	V
42.20	U
…	…

Sorted SNPs



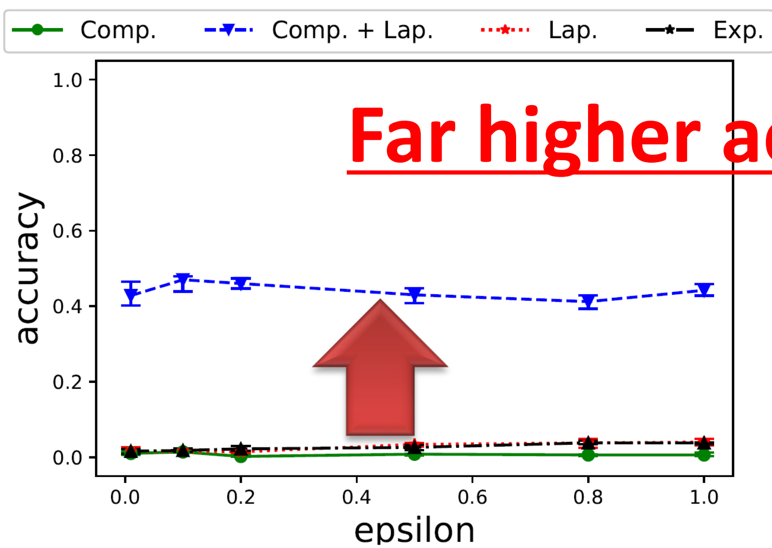
Result

[RECOMB Genome Security Workshop, 2022 (JCB 2023)]

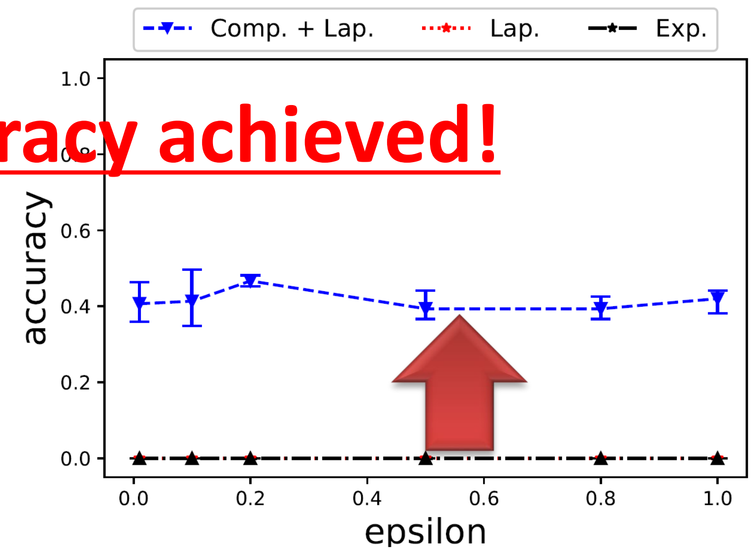
T. Shibuya

The top-10 significant SNPs

Far higher accuracy achieved!



Simulated Data (#SNPs=500)



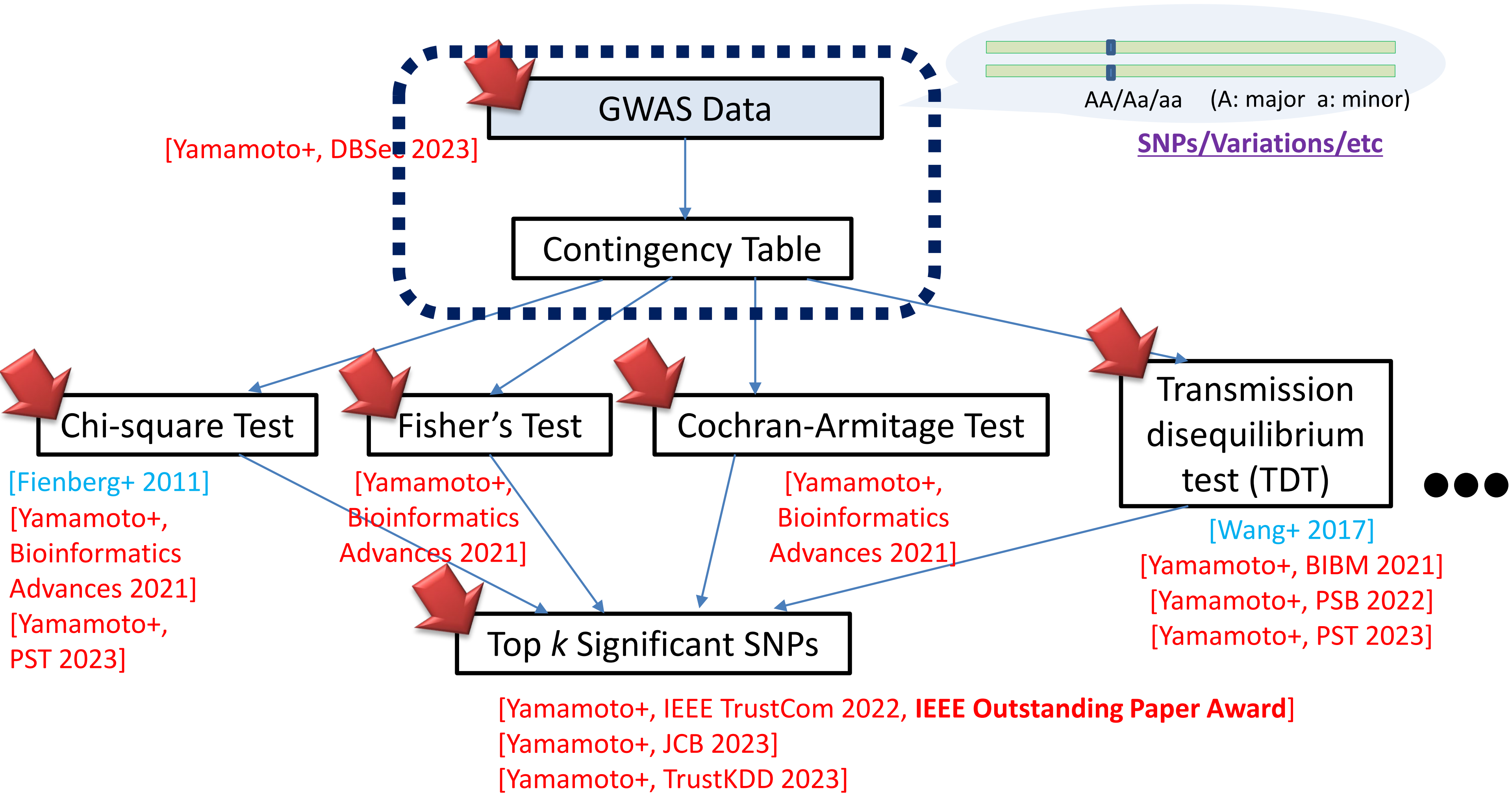
Simulated Data (#SNPs=25,000)

Running time (sec)

Mechanism	#SNPs=500	#SNPs=25,000
Ours (Comp+Lap)	2.96	7.9×10^{3}
Compressive	6.52	- (Takes too much time)
Laplace	2.9×10^{-4}	5.6×10^{-3}
Exponential	1.6×10^{-3}	7.8×10^{-2}

DP mechanisms for GWAS

■ Differentially Private Mechanism Design for GWAS



Local Differential Privacy

- **Local differential privacy** [Kasviswanathan et al., 2008]
  - Add noise to all the data labels 'locally'
    - No one (except for the data owner) can see the original data, while we can do any analysis on the published noise-added data
- **Strategies**
  - Ordinary DP mechanisms for numerical data
    - e.g., Laplace mechanism
  - Random response for label data [Warner+ 65]
    - Changing labels probabilistically
      - e.g., flipping 0/1 probabilisitically for 0/1 data

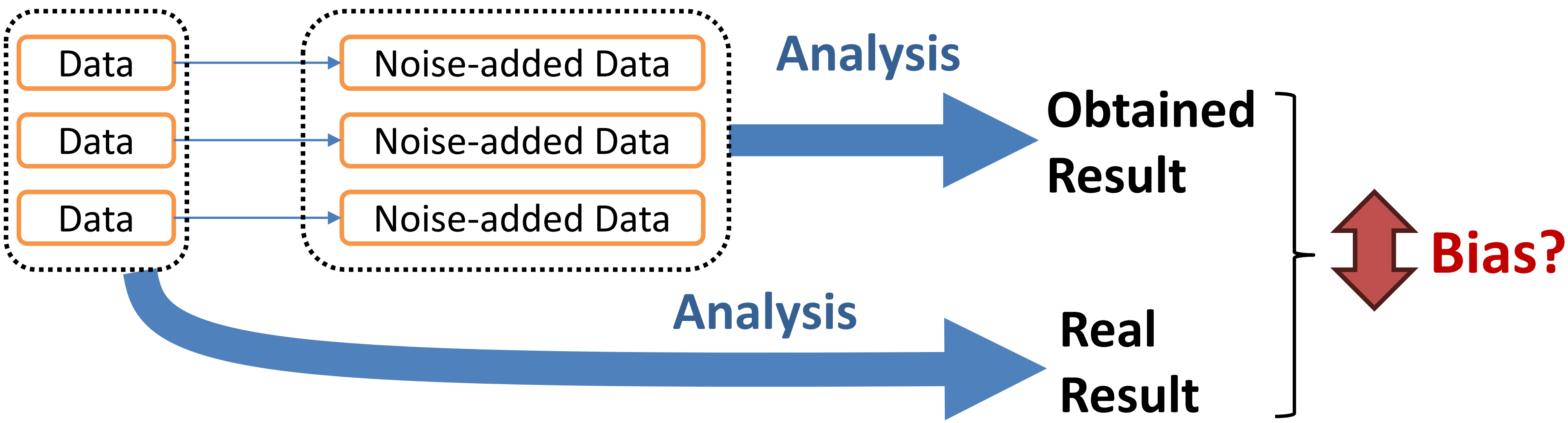
	0	1
0	$1-\alpha$	α
1		$1-\alpha$

$(\alpha = \frac{1}{e^{\varepsilon}+1})$

Distortion matrix

Analysis Bias by Local Differential Privacy

- Local differential privacy schemes could cause biases
- Debiasing methods
  - EM-algorithm for random response
    - RAPPOR [Erlingsson+ 14]
    - GWAS contingency table [Yamamoto+ 23]
  - Debiasing polynomial functions for Laplace noise
    - *k*-star counting on graphs [Hillebrand+ 23]



More Accurate Strategies for Contingency Tables

[HEALTHINF 2023]

T. Shibuya

- Consider attribute pair as a single attribute to reduce noise
- EM algorithm to improve accuracy
  - Compute $\mathrm{argmax}_{P,Q,R,S}\ \mathrm{Prob}(P', Q', R', S' | P, Q, R, S)$

	Case	Control
Major	P	Q
Minor	R	S

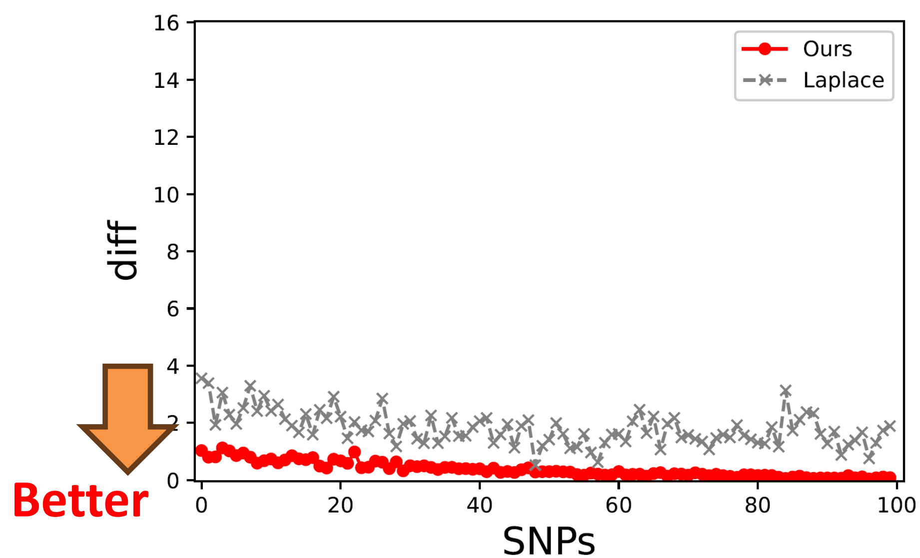
Randomized Response ↓ ↑ Expectation Maximization

	Case	Control
Major	P´	Q´
Minor	R´	S´

	P	Q	R	S
P	$1-3\alpha$	α	α	α
Q		$1-3\alpha$	α	α
R			$1-3\alpha$	α
S				$1-3\alpha$

$\left(\alpha = \frac{1}{e^{\varepsilon}+3}\right)$

Distortion Matrix $\mathrm{Prob}(X \to Y)$

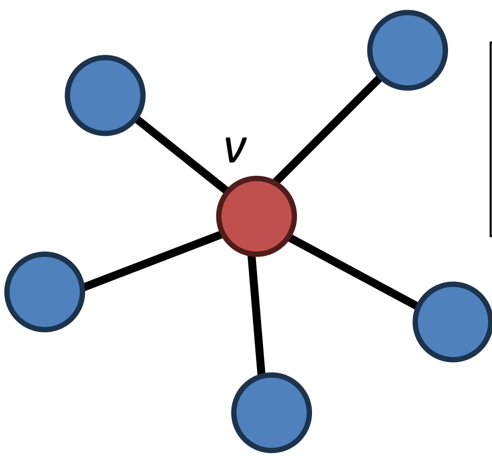


Cochran-Armitage Trend Test Accuracies

Local Differential Privacy on Graph

[Hillebrand+, KDD 2023]

- Assume a graph where
  - Each vertex has its "sensitive" adjacency list
- Problem
  - Number of k-stars in graph
- Strategy
  - Each vertex provide its Laplace noise-added degree
  - Compute number of k-stars based on the reported degrees



Send its degree with noise a=(5 + noise)

Estimated #k-star is: $\Gamma(a+1)/\Gamma(a-k+1)\Gamma(k+1)$

e.g. 6.31 → *e.g.*Estimated #*k*-star is: 24.1

Estimation

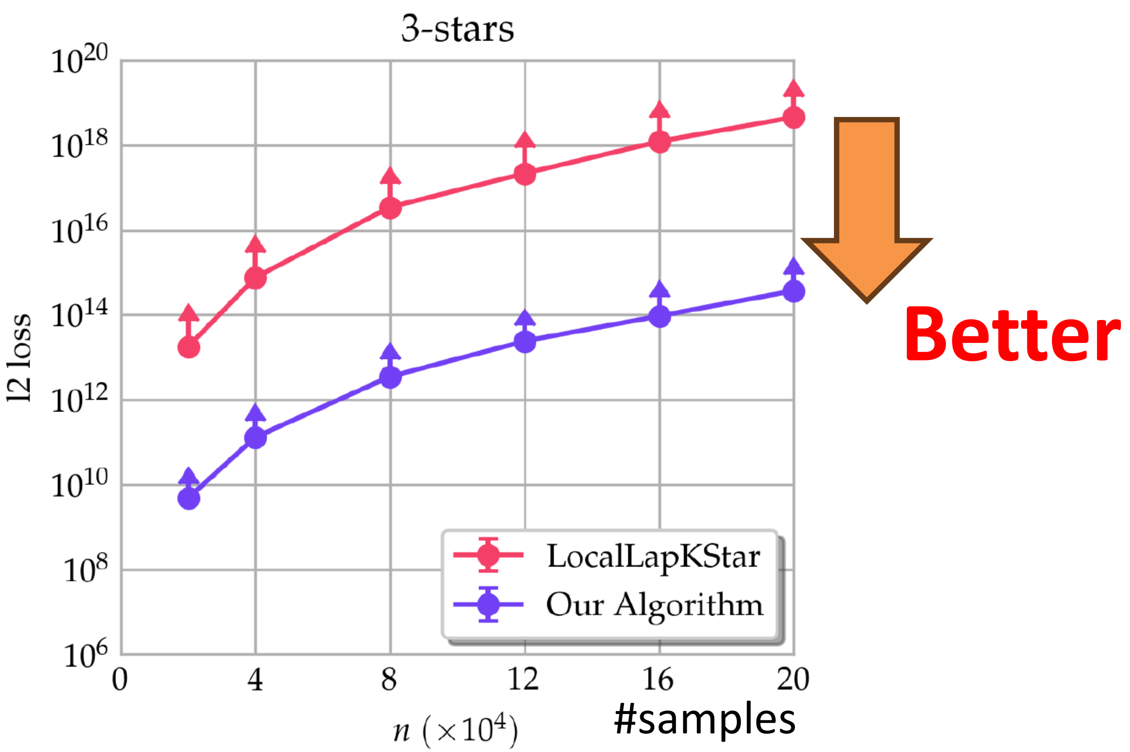
Large difference!!

Actual #3-star around *v* is $\binom{5}{3} = 10$

Debiasing Polynomial Effect of Laplacian Noise [Hillebrand+, KDD 2023]

- Tomasz's Theorem [Tomasz+ 10]
  - The expected value of Laplace noise-added x^r
    - $E\left((x + Lap(x,b))^r\right) = \sum_{k=0}^{\lfloor r/2 \rfloor} \frac{\Gamma(r+1)}{\Gamma(r-2k+1)} b^{2k} x^{r-2k}$
- Experiment
  - Estimating #3-stars on IMDB datasets [Leskovec+ 14]
    - (896,308 nodes/ 57,064,358 edges)

L2 Loss



#samples

Assigning Different Privacy Budgets to Many Attributes [ISCC 2023]

Problem of finding the 'optimal' distortion table

- Objective
  - Minimize the entire privacy budget
- Parameters
  - $O(|\Sigma|^{2l})$ flip probabilities between all pair of $O(|\Sigma|^{l})$ label strings
    - Σ: #label (= alphabet), l: #attributes (=string length)
- Constraints
  - Given different privacy budgets for different attributes
  - 'Reasonable' flip probabilities

Different privacy requirements

ID	A	B	C	D	E	F	G
1	1	0	0	1	2	0	1
2	1	1	1	0	2	0	0
3	0	2	1	0	1	0	1
4	2	0	0	1	0	1	2
5	1	1	0	0	1	1	0

Linear Programming

Faster Heuristic

Objective

- Consider each data as a string $S_i \in |\Sigma|^l$
- Minimize the entire privacy budget
  - i.e., $\max_{ijkl} (p_{ij}/p_{kl}) \quad (i \neq j, k \neq l)$

	S_1=000	S_2=001	S_3=010	S_4=011	S_5=100	S_6=101	S_7=110	S_8=111
S_1=000	p_{11} No edit	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}	All edited p_{18}
S_2=001		p_{22}	p_{23}	p_{24}	p_{25}	p_{26}	p_{27}	p_{28}
S_3=010			p_{33}	p_{34}	p_{35}	p_{36}	p_{37}	p_{38}
S_4=011				p_{44}	p_{45}	p_{46}	p_{47}	p_{48}
S_5=100					p_{55}	p_{56}	p_{57}	p_{58}
S_6=101						p_{66}	p_{67}	p_{68}
S_7=110							p_{77}	p_{78}
S_8=111								p_{88}

Distortion matrix for 3 binary attribute data

Budget Constraint for Each Attribute

- Privacy budget ε_i for each attribute i is given
  - e.g. $\sum_* \mathrm{P}(100 \to *0*) / \sum_* \mathrm{P}(100 \to *1*) \le e^{\varepsilon_2}$
    - The same for 000/001/101

	S_1=000	S_2=001	S_3=010	S_4=011	S_5=100	S_6=101	S_7=110	S_8=111
S_1=000	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}	p_{18}
S_2=001		p_{22}	p_{23}	p_{24}	p_{25}	p_{26}	p_{27}	p_{28}
S_3=010			p_{33}	p_{34}	p_{35}	p_{36}	p_{37}	p_{38}
S_4=011				p_{44}	p_{45}	p_{46}	p_{47}	p_{48}
S_5=100					p_{55}	p_{56}	p_{57}	p_{58}
S_6=101						p_{66}	p_{67}	p_{68}
S_7=110							p_{77}	p_{78}
S_8=111								p_{88}

Distortion matrix for 3 binary attribute data

'Reasonable' Edit Transition Probability Constraint

- Larger edit (response) should be rarer
  - e.g., $P(010 \rightarrow 011) \geqq P(010 \rightarrow 001)$
    - which corresponds to edit transition $010 \rightarrow 011 \rightarrow 001$

	S_1=000	S_2=001	S_3=010	S_4=011	S_5=100	S_6=101	S_7=110	S_8=111
S_1=000	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}	p_{18}
S_2=001		p_{22}	p_{23}	p_{24}	p_{25}	p_{26}	p_{27}	p_{28}
S_3=010			p_{33}	p_{34}	p_{35}	p_{36}	p_{37}	p_{38}
S_4=011				p_{44}	p_{45}	p_{46}	p_{47}	p_{48}
S_5=100					p_{55}	p_{56}	p_{57}	p_{58}
S_6=101						p_{66}	p_{67}	p_{68}
S_7=110							p_{77}	p_{78}
S_8=111								p_{88}

Distortion matrix for 3 binary attribute data

Reducing #Parameters by Utilizing Symmetry in Distortion Matrix

T. Shibuya

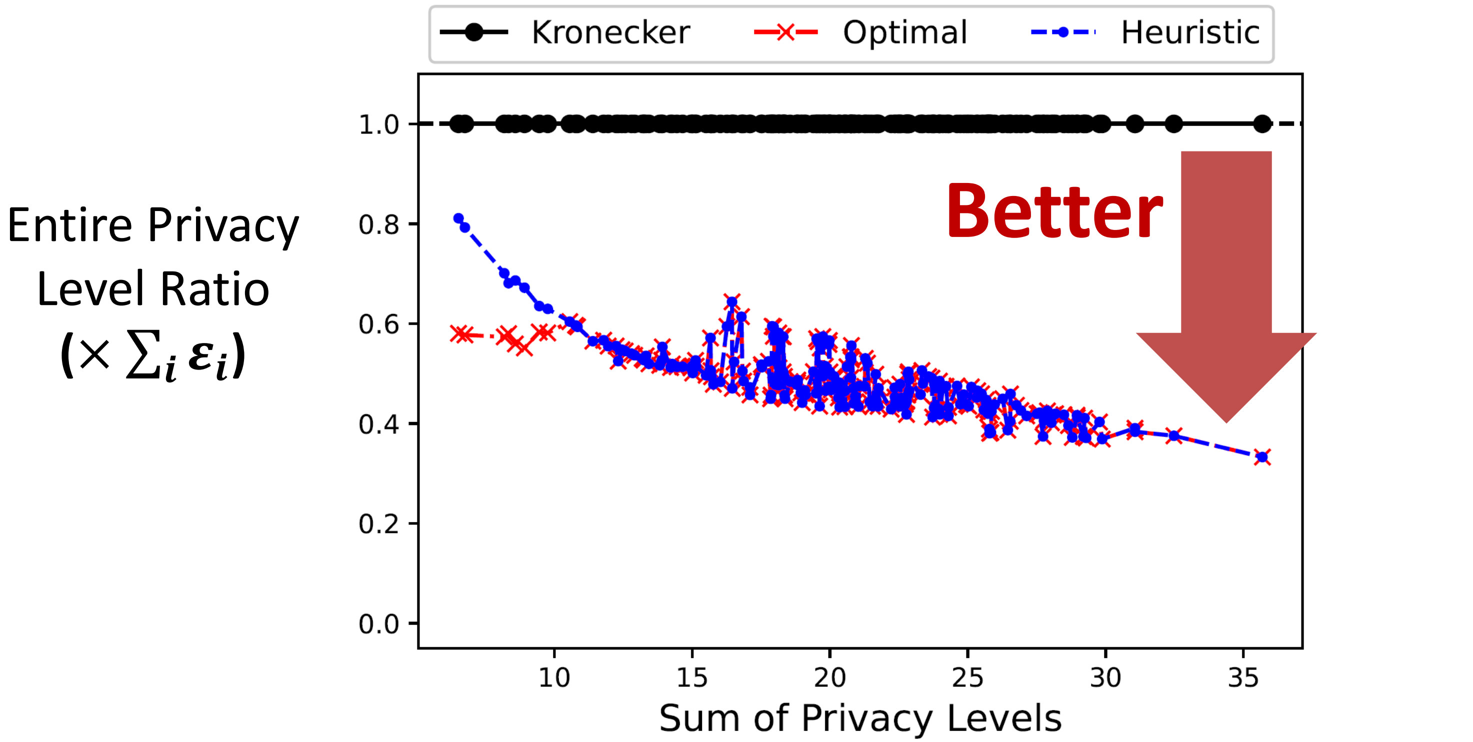
- The same editing probabilities for the same set of attributes
  - regardless of labels
  - e.g. P(000→101)=P(001→100)=P(010→111)=P(011→110)

	S_1=000	S_2=001	S_3=010	S_4=011	S_5=100	S_6=101	S_7=110	S_8=111
S_1=000	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}	p_{18}
S_2=001		p_{22}	p_{23}	p_{24}	p_{25}	p_{26}	p_{27}	p_{28}
S_3=010			p_{33}	p_{34}	p_{35}	p_{36}	p_{37}	p_{38}
S_4=011				p_{44}	p_{45}	p_{46}	p_{47}	p_{48}
S_5=100					p_{55}	p_{56}	p_{57}	p_{58}
S_6=101						p_{66}	p_{67}	p_{68}
S_7=110							p_{77}	p_{78}
S_8=111								p_{88}

Distortion matrix for 3 binary attribute data

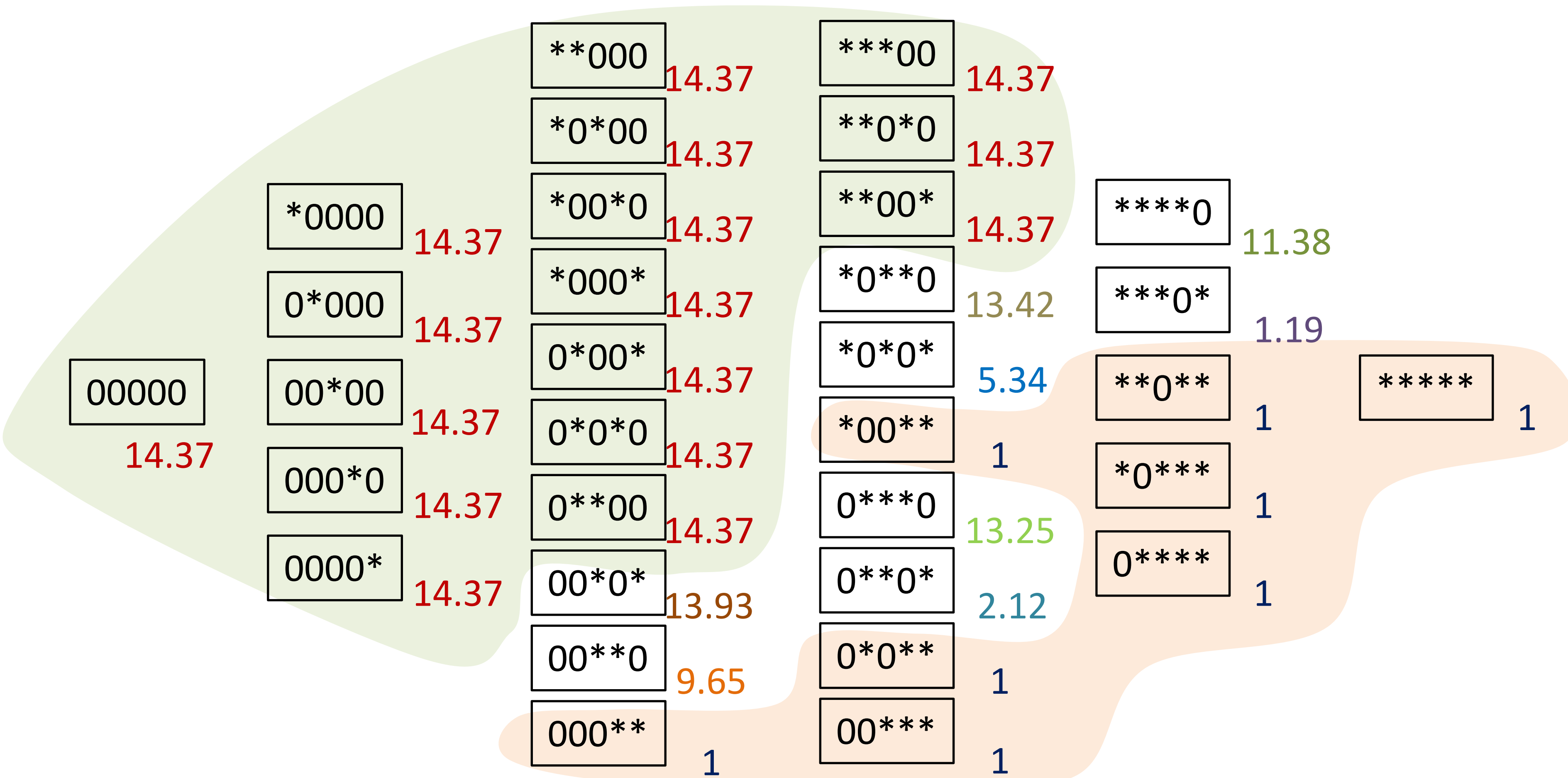
Experimental Result

- Entire privacy levels of the optimal distortion matrices
  - on randomly generated attribute privacy budgets
    - 5 attributes, $|\Sigma| = 5$, $1 \leq \varepsilon_i \leq 8$, 200 sets
- Our heuristic also achieves near-optimal privacy level



An Example of the Optimal Solution

- Example Settings:
  - 5 attributes, $|\Sigma| = 5$, $\varepsilon_1 = 0.1, \varepsilon_2 = 0.2, \varepsilon_3 = 0.5, \varepsilon_4 = 0.7, \varepsilon_5 = 2.0$
- Optimal entire privacy level: ε =2.67 (=ln 14.37)
  - Better than the simple strategy [Wang+ 16] where $\boldsymbol{\varepsilon} = \sum_i \boldsymbol{\varepsilon_i} = 3.5$



k-Anonymization: A Yet Another Privacy Preservation Technique

T. Shibuya

[Sweeney 2002]

- To reduce risk of being identified
  - 85% of the US citizens can be identified only by (birthdate/ZipCode/Sex) information [Sweeney 2002]

Name	Birthdate	Zip Code	Sex	Information
Alex Tokyo	19990123	108-8639	Male	…
Robert Kyoto	19990711	153-8902	Male	…

k=1

k=1

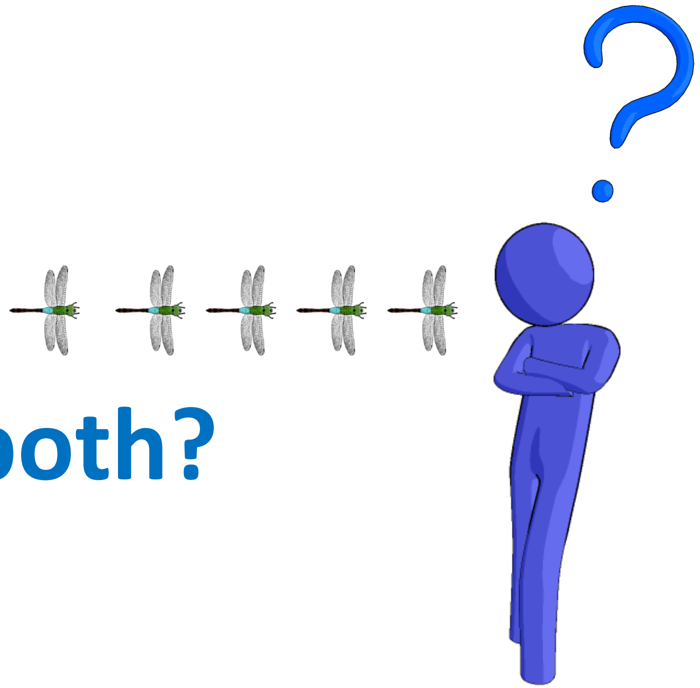
Name	Birthdate	Zip Code	Sex	Information
PB924CD	1999****	1**-8***	Male	…
AR325HB	1999****	1**-8***	Male	…

k=2

2-Anonymization

k-Anonymization and Differential Privacy

- k−anonymization does not satisfy the differential privacy
- Differential privacy does not satisfy the k−anonymization
  - Noise added data can collide with the existing data in coincidence
    - It could cause a problem of false accusation



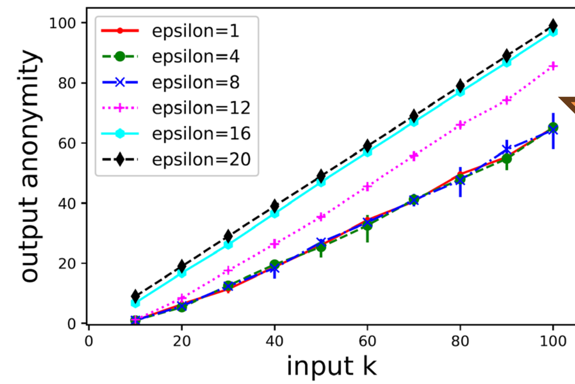
Can we satisfy both?

Strategies

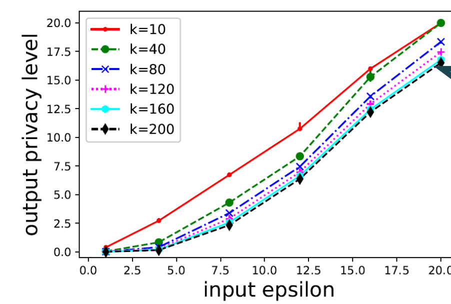
- Naïve algorithm (kN+DP)
  - k-anonymization BEFORE differential private anonymization
    - k-anonymity not satisfied
- Naïve algorithm (DP+kN)
  - k-anonymization AFTER differential private anonymization
    - Both anonymity satisfied, but less accurate
      - Due to the **too ‘high’** privacy level
- Our algorithm ((ε, k)–anonymization)
  - $k'(k' < k)$–anonymization first
    - To prevent accuracy loss in the final k-anonymization
  - Then, Differential private anonymization
  - k-anonymization, finally
    - Satisfies both anonymity, keeping accuracy

Experimental Results

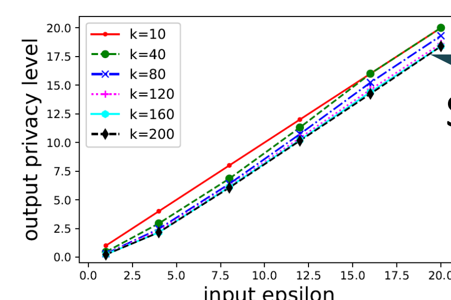
- **Data**
  - 1,512, 673 entries, J−MIMO Medical Record 2021−3.
- **Results**
  - kN+DP
    - k-anonymity not satisfied
  - DP+kN
    - Privacy level increases unintentionally
      - which causes substantial loss of accuracies
  - (ε, k)−anonymization
    - Very accurate, satisfying both properties



k-anonymity loss of kN+DP



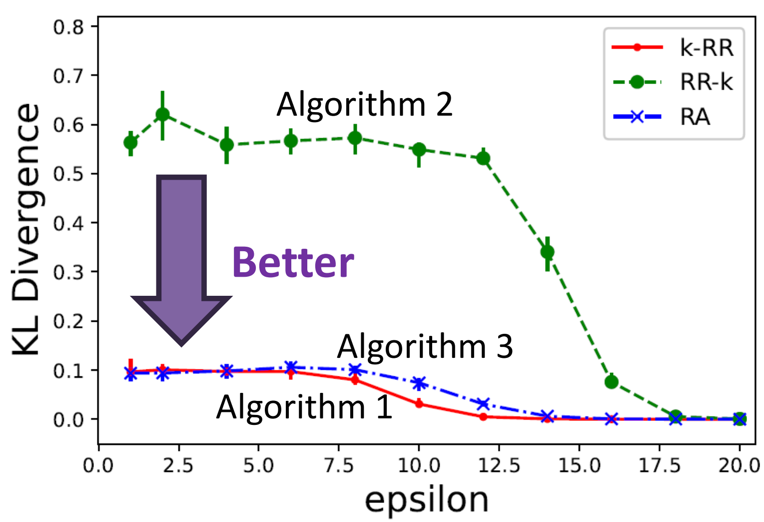
DP+kN



Smaller

(ε, k)-anonymization

Unintentional increase of privacy levels



Accuracies of the algorithms

Summary / Future Work

- Differentially private methods for biomedical data
  - GWAS statistics publication
  - Post-processing for local differential privacy
  - Multiple attribute publication
  - *k*-anonymization and differential privacy
- For the CPM community ☺
  - A string = A set of multiple attributes
  - We could consider differential privacy on many CPM problem (preferably on sensitive data)
    - How to reduce noise (to a reasonable level)
    - How to debias

Acknowledgements

Quentin Hillebrand

Akito Yamamoto

Prof. Vorapong Suppakitpaisarn

Thank you very much!