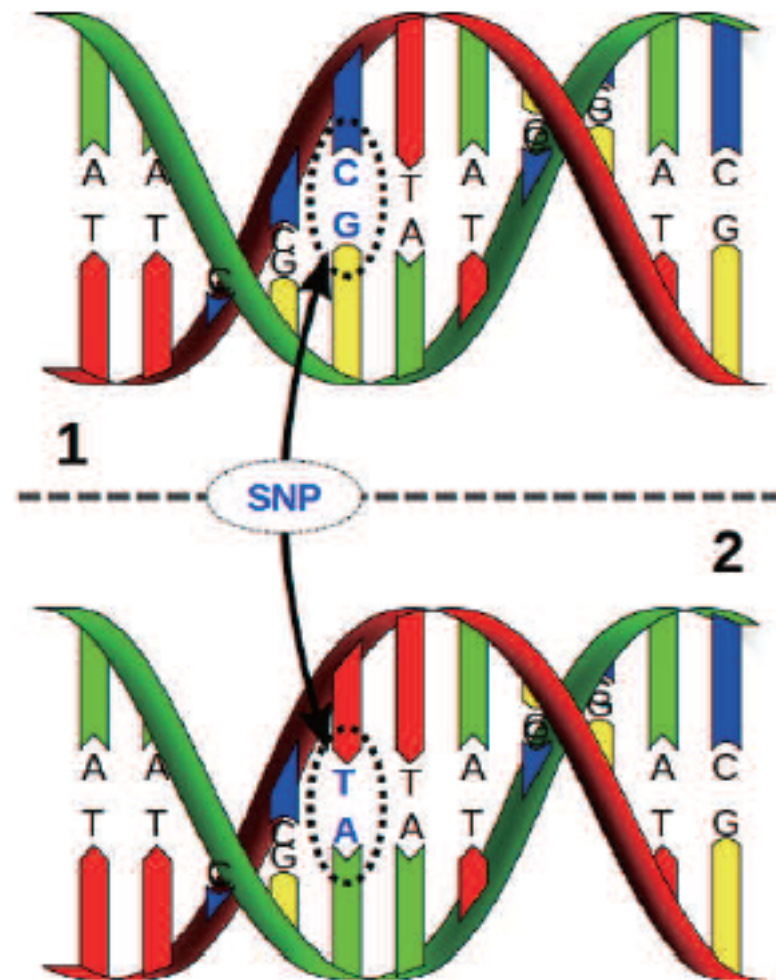


ゲノムプライバシーの保護と 個別化医療への展開

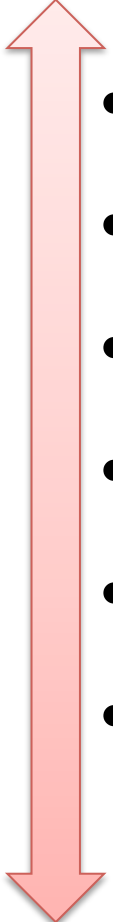
筑波CS/JST CREST 佐久間 淳

個人ゲノム



個人ゲノムでわかること

あたる

- 
- 耳あかのタイプ、乳糖不耐性
 - 2型糖尿病の罹患リスク
 - 祖先はどの大陸から来たか？
 - アルコールを飲むと赤くなるか
 - 生活習慣病の罹患リスク
 - 性格、IQ、能力、

微妙(後天的要因が大きい)

個人ゲノムに基づく解析

- 類似性に基づく鑑定
 - 本人鑑定、父子鑑定、親戚鑑定
- ゲノム疫学
 - 疾患関連遺伝子探索
 - 疾患リスクモデル予測
- 個別化医療/出生前診断
 - 疾患リスク予測に基づく予防医療
 - 薬剤感受性評価

疾患関連遺伝子探索

- 分割表 (e.g, 心筋梗塞)

	対立遺伝子 AA, Aa	対立遺伝子 aa	合計
ケース	a	b	$a + b$
コントロール	c	d	$c + d$
合計	$a + c$	$b + d$	n

- 検定
 - EX. 帰無仮説「対象疾患と、このSNPは独立」についてカイ二乗検定
- ゲノムワイド相関解析
 - 全SNPについて対象疾患との関連の強さを網羅的に検定

疾患リスク予測

遺伝的容易由来のリスク 臨床的要因由来のリスク

$$\text{risk} = \mathbf{w}^T \mathbf{x} = \underbrace{w_1^G x_1^G + w_2^G x_2^G + \dots}_{\text{遺伝的容易由来のリスク}} + \underbrace{w_1^C x_1^C + w_2^C x_2^C + \dots}_{\text{臨床的要因由来のリスク}}$$

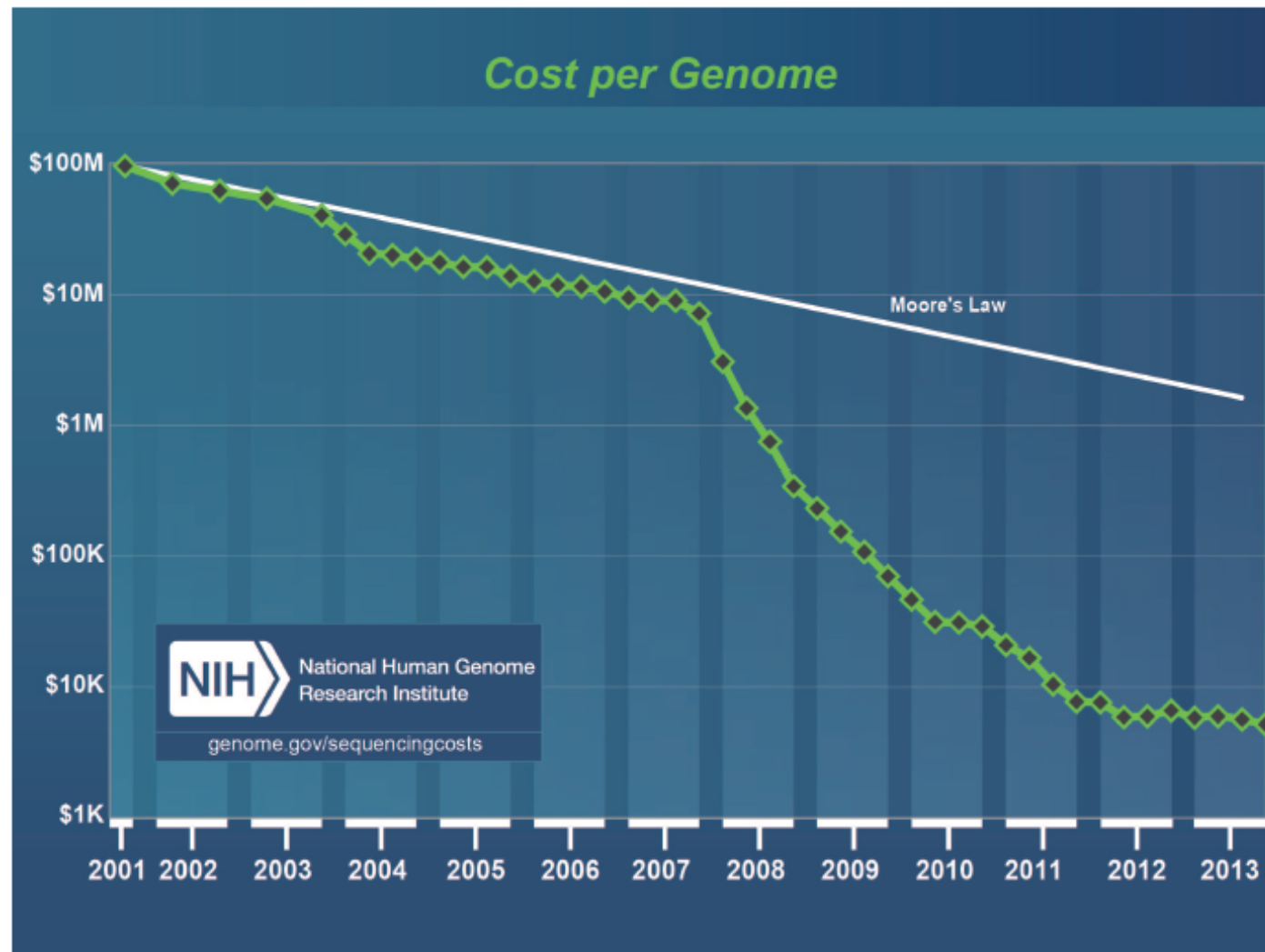
Presence of SNP1 A clinical factor (e.g. age)
effect of SNP1 on the risk of developing DT effect of the factor on the risk of developing DT

- 遺伝的要因
 - E.g., SNP nの対立遺伝子がAAであるか否か
- 臨床・環境的要因
 - 年齢、性別
 - 飲酒歴
 - 喫煙歴, etc.

個人ゲノムにまつわるfact

- Good
 - － 個人の様々な体質を表現
 - － 解析コストが劇的に低下
 - － 生涯変化しない

個人ゲノムの解析コスト



<http://www.genome.gov/sequencingcosts/>

個人ゲノムにまつわるfact

- Good

- 個人の様々な体質を表現
- 解析コストが劇的に低下
- 生涯変化しない

- Bad?

- 強力な識別子であり、かつ、属性情報 (プライバシー)
- 詐取が容易 (オレオレゲノム)
- 個人ゲノムがもたらすリスクは現時点で不明
- 遺伝的差別の可能性

個人ゲノム=強力な識別子+詳細な個人情報

- 個人情報

ID	名前	住所	年齢	病院名	疾病	薬
0A99934	筑波太郎	茨城県	32	A病院	腰痛	XXX
1B92372	統計花子	東京都	46	B医院	鼻炎	YYY
3C88892	匿名幸	京都府	42	Cクリニック	水虫	ZZZ

識別子

属性情報

- 個人ゲノム

ID	名前	ABCC8	APOEA1	ABCA1	APOE	AKAP10
0A99934	筑波太郎	AG	GG	CC	CT	GA
1B92372	統計花子	GG	GG	CT	CC	AA
3C88892	匿名幸	GG	GG	CC	CC	AA

識別子

識別子: 本人鑑別に使えるぐらい強力
属性情報: そのものが形質を表現

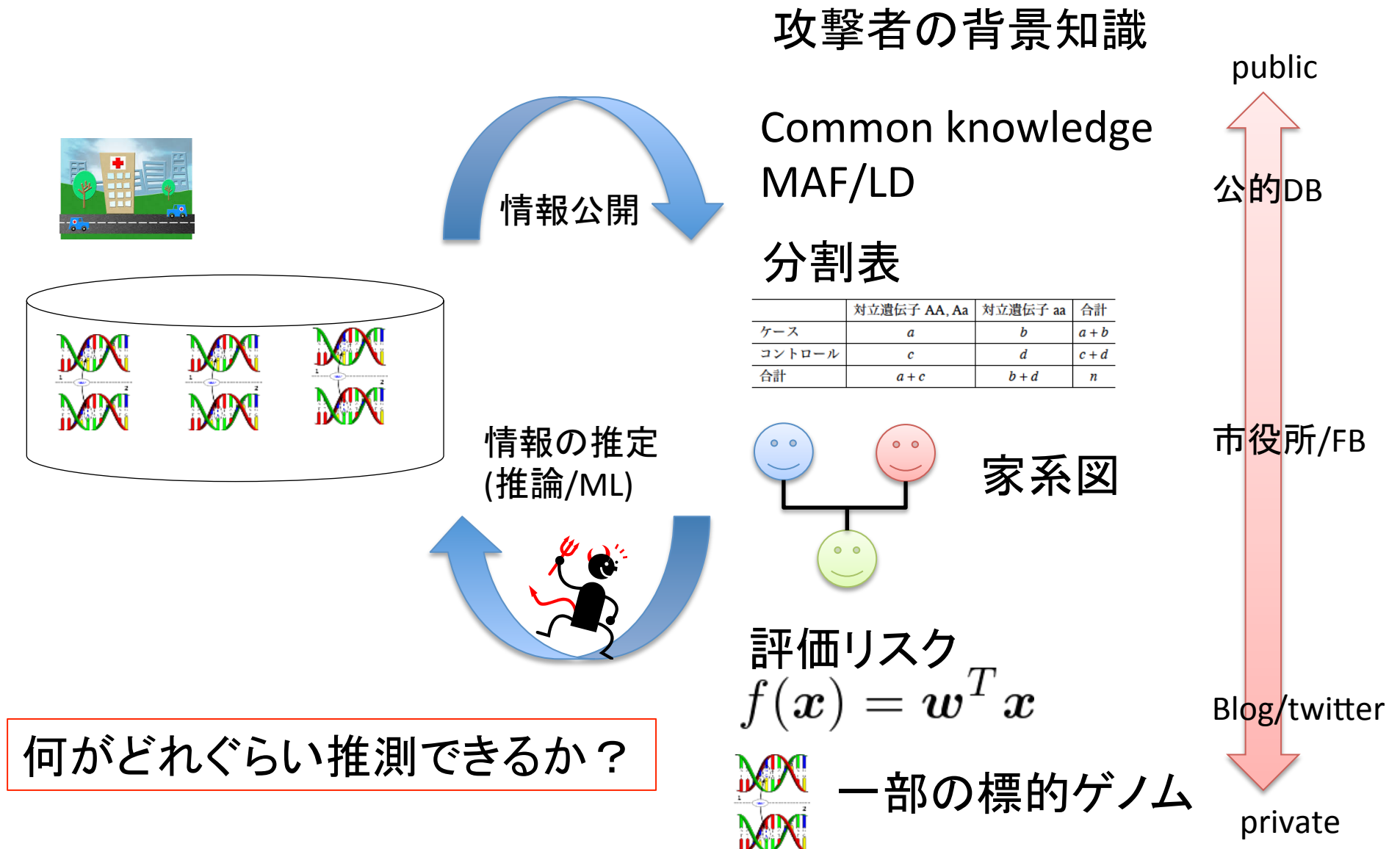
予期できないリスク

- personal genome project (PGP)
 - ボランティアベースでの個人ゲノムデータの収集・提供
 - 将来において”匿名性は保証されない”ことへの同意(open consent)が前提
- PGPが提示する“リスク”
- 父系やその他血縁関係の推定
- 雇用、保険、金融サービスに影響する統計的証拠の推定
- 被験者やその親戚の犯罪性向の推定
- 被験者に相当するDNAを合成し、これを犯罪目的で(誰かに)埋め込む
- 効果的な治療法が知られていない疾患の罹患リスク開示

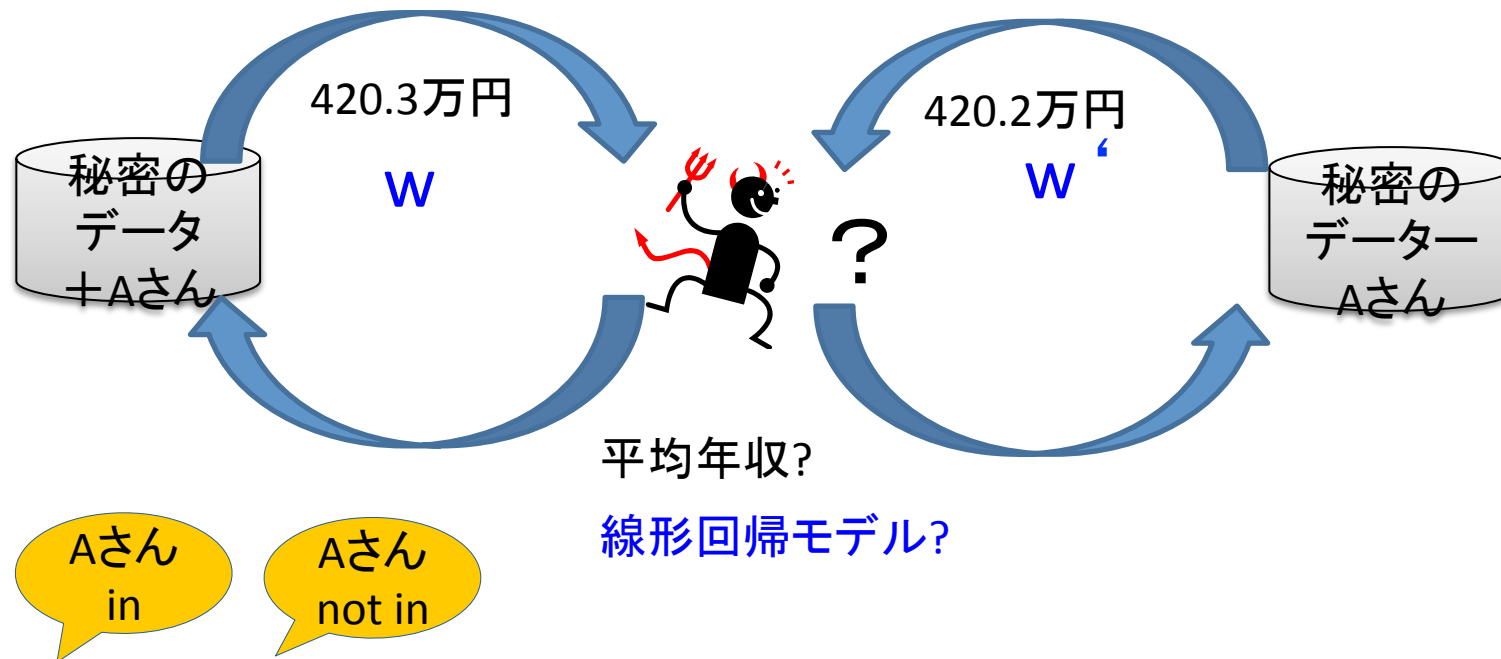
社会上の影響

- 遺伝子差別
 - 遺伝子による...就職差別,結婚差別,保険差別
 - 米国は医療保険や採用の遺伝的差別法律により既に禁じている
- 教育への影響:あなたの子供が将来...
 - プロ野球選手になれる可能性はx%
 - 大学教授になれる可能性はy%
- ゲノムクライム
 - ゲノムなりすまし
 - ゲノム脅迫?「2ちゃんにおまへのゲノムのせるぞ!」
- 遺伝子に基づく広告?

機械学習の立ち位置: 統計的推論による攻撃

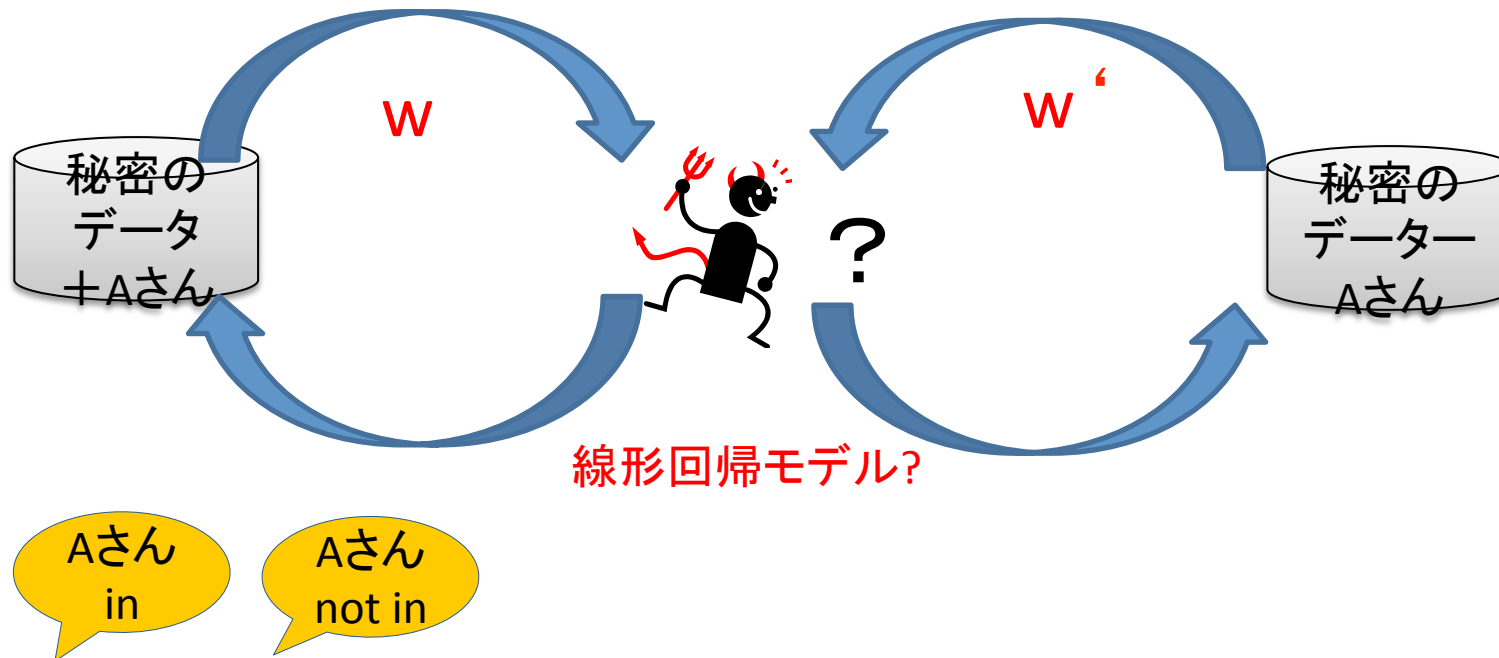


統計DBクエリのSemantic security



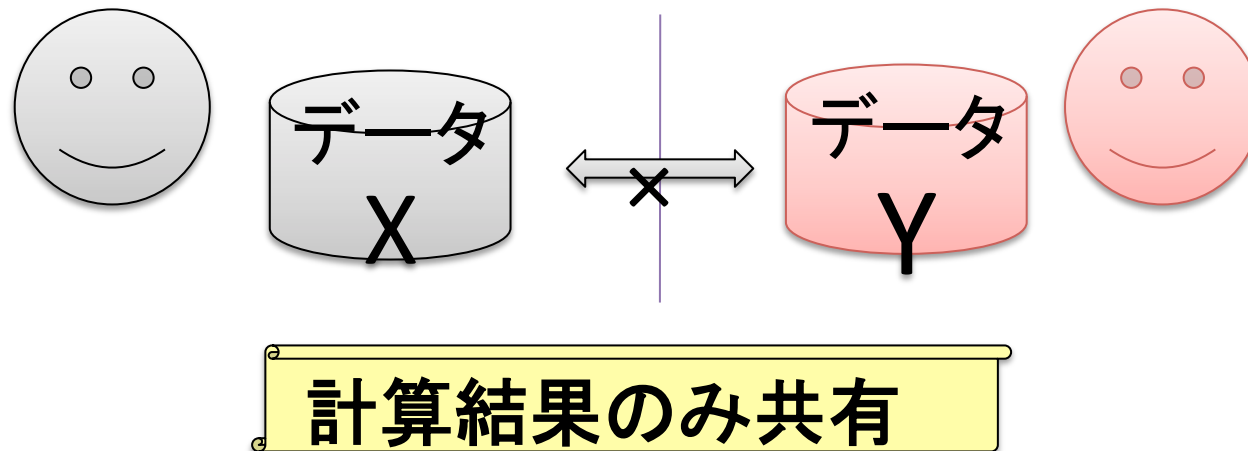
- $f(D+A)$ と $f(D)$ がそんなに変わらなければ、 $f(D+A)$ の開示はAさんのプライバシーを侵害していない (ことにしよう！)

統計DBクエリのSemantic security



- $f(D+A)$ と $f(D)$ がそんなに変わらなければ、 $f(D+A)$ の開示はAさんのプライバシーを侵害していない (ことにしよう！)

プライバシー保護の技術:秘密計算



- AはXをBに開示したくない, BはYをAに開示したくない
- ただし (X, Y) について $f(X, Y)$ を計算し結果のみ知りたい
- 信頼できる第三者(TTP)なしにこれを実現したい
- アプローチ: 暗号理論に基づく安全な分散計算

Computation over encrypted values

Holomorphic encryption



m_1, m_2 :secret numbers

$c_1 = \text{Enc}(m_1)$: encryption of m_1  c_1

$c_2 = \text{Enc}(m_2)$: encryption of m_2  c_2

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

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

疾患リスク予測

遺伝的容易由来のリスク 臨床的要因由来のリスク

$$\text{risk} = \mathbf{w}^T \mathbf{x} = \underbrace{w_1^G x_1^G + w_2^G x_2^G + \dots}_{\text{Genetic risk}} + \underbrace{w_1^C x_1^C + w_2^C x_2^C + \dots}_{\text{Clinical risk}}$$

Presence of SNP1 A clinical factor (e.g. age)
effect of SNP1 on the risk of developing DT effect of the factor on the risk of developing DT

- 遺伝的要因
 - E.g., SNP nの対立遺伝子がAAであるか否か
- 臨床・環境的要因
 - 年齢、性別
 - 飲酒歴
 - 喫煙歴, etc.

プライバシーを保護したリスク予測

- 疾患リスク予測

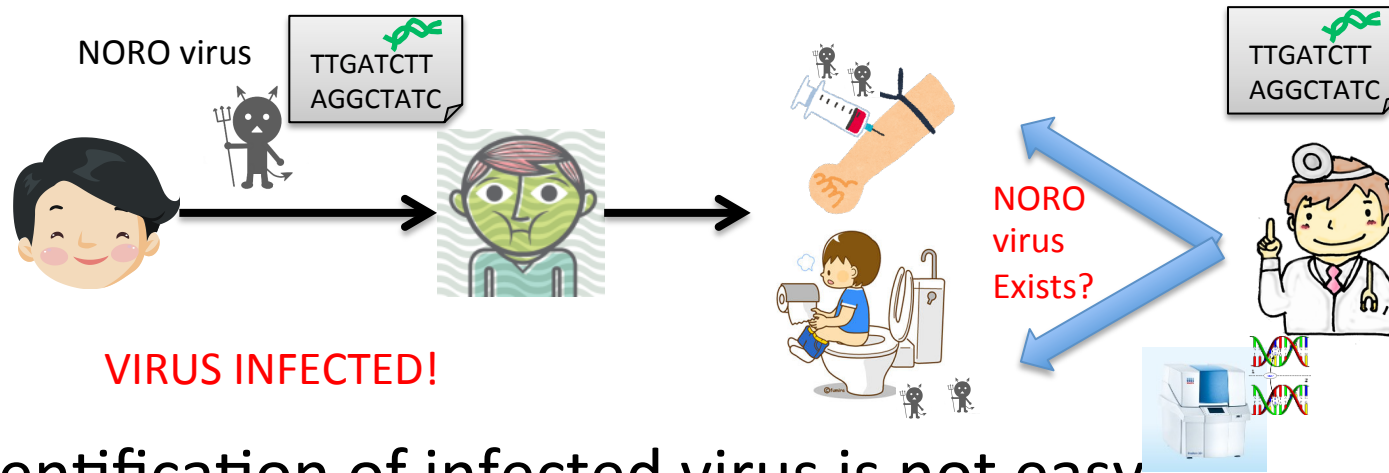
$$\boldsymbol{w}^T \boldsymbol{x} = \sum_i w_i^G x_i^G + \sum_i w_i^C x_i^C$$

- 準同型暗号上の疾患リスク予測

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

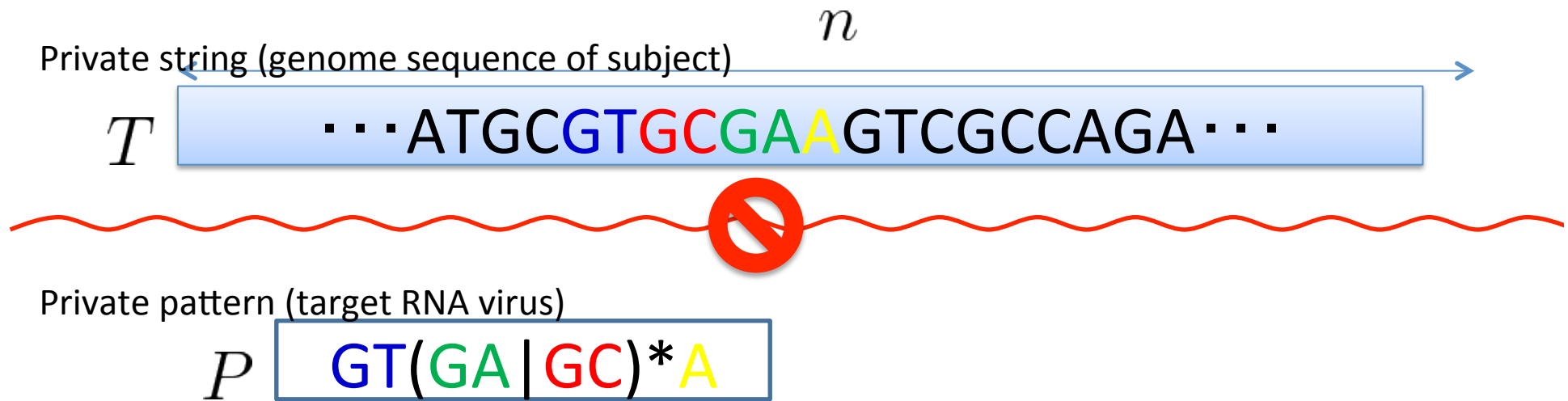
Virus detection from RNA

[Sasakawa+WPES2014] Oblivious Evaluation of Non-deterministic Finite Automata with Application to Privacy-Preserving Virus Genome Detection]



- Identification of infected virus is not easy
 - Electronic microscope
 - Cloning
- Virus RNA detection from patients samples (blood, faecals) by next generation sequencer
 - Quick and relatively inexpensive
 - Privacy concern remains
 - Personal genome is sequenced together with virus RNA
 - Bio terrorism

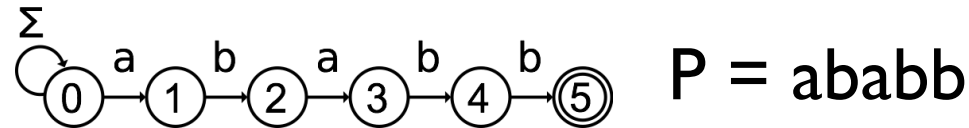
Privacy-preserving virus detection with string matching



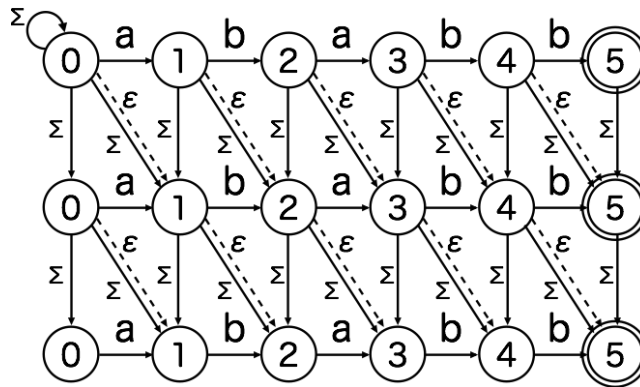
- Find presence of private pattern (virus) P in private text T (genome sequence)
- Learn presence only, nothing else

String Search via NFA Evaluation

Linear NFA:
Exact search

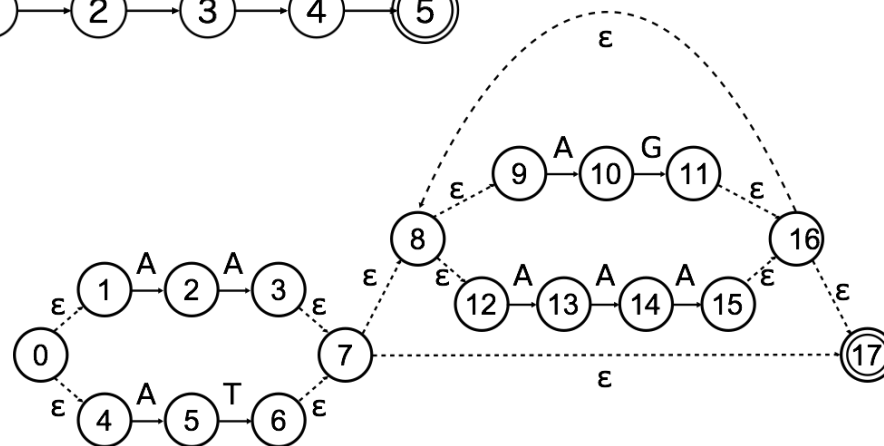


Ukkonen NFA:
Approximate search



$P = ababb$
2-differences are
permitted

Thompson NFA:
Regular expression search



$P = (AA|AT)((AG|AAA)^*)$

Matrix Representation of NFA

	0	1	2	3	4
b	0	1	2	3	4
0	1	1	1	1	1
1	1	1	0	1	1
2	1	1	0	1	1
3	1	1	1	1	1
4	1	1	1	1	1

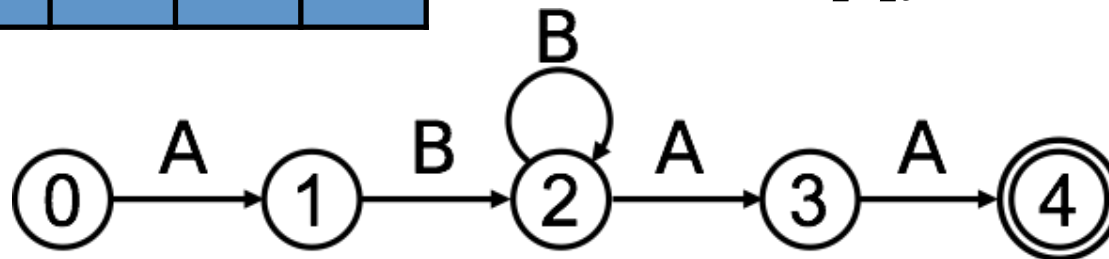
As a preprocessing, the algorithm constructs adj-matrix of NFA for each edge label.

- If NFA has the edge from j to k labeled by character c

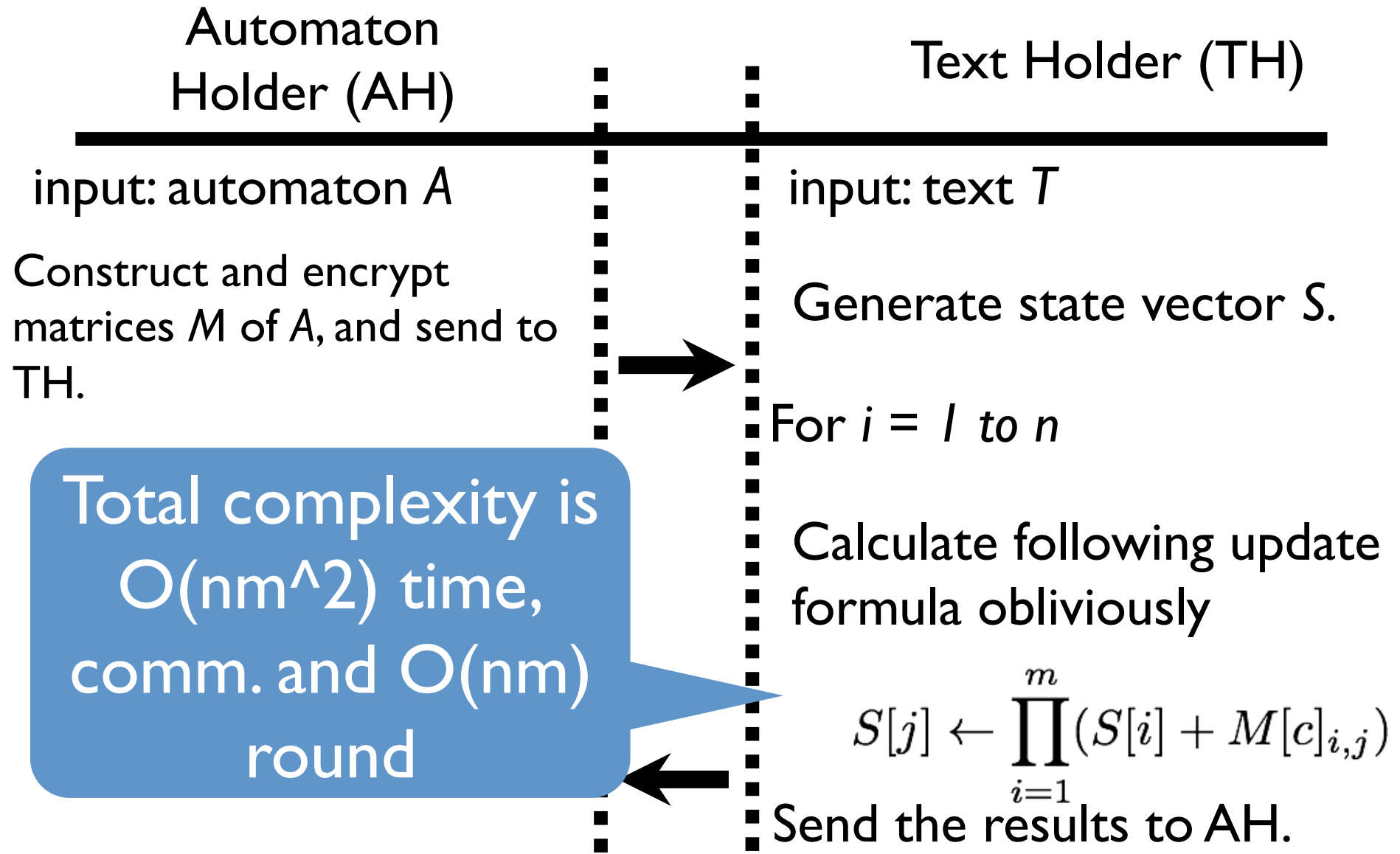
$$- M[c]_{j,k} = 0$$

- Otherwise

$$- M[c]_{j,k} = 1$$



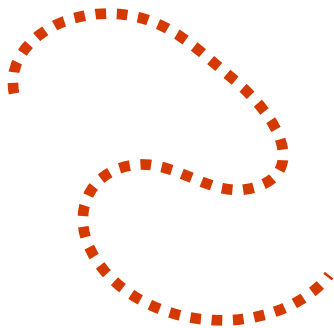
ONE Protocol



m : the size of automaton A , n : the length of text T

RNA Virus Detection

RNA Virus:



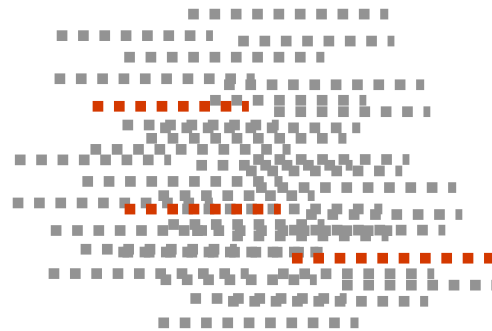
E.g. **Norovirus**

Approximately 6,500 base pairs (4-letter alphabet). Can be read in either direction

☞ $\approx 13,000$ letters

Mutations (letter substitutions) can occur.

Sequencing Data:

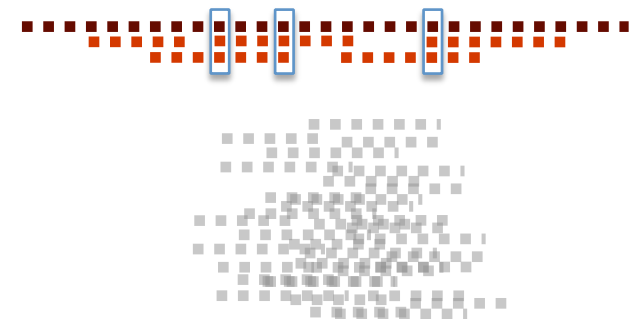


Fragments of RNA (“**short read**”):
each sample ≈ 150 bases $\times$ $3 \sim 6$ M

Includes RNA from many different organisms (metagenome), not just virus or host.

If infected: $\approx 30,000$ reads belong to the virus ($\approx 1.2\%$).

Detecting the virus:



Map short reads to a **reference genome** for the virus: substring matching task allowing for insertion/deletion/substitution.

If the virus is present in the sample, each position of the reference genome will be covered at least a certain number of times by mapped short reads.

Oblivious RNA Virus Detection

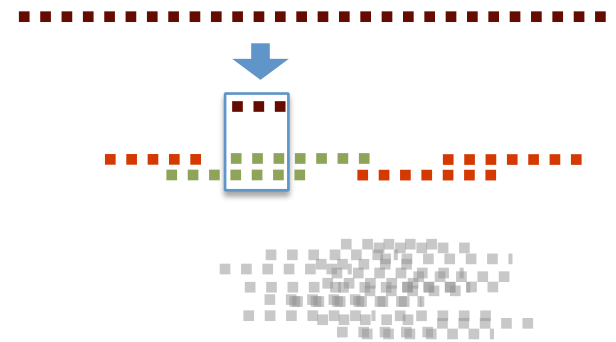
Mapping all short reads to the reference genome would be too expensive in a privacy-preserving context.

Instead, we select a highly-preserved substring (31 letters) from the reference genome and use it to test a sample of short reads (30,000 or about 0.5% of all reads) using our **oblivious k-mismatch algorithm** (ONE).

👉 **Worst-case probability of false negative < 0.074%**

Results on 3 infected samples and 1 control:

Sample	Total reads	Exact Matches	k = 1	k = 2
Norovirus samples				
ERX202484	2957446	0 (0)	474 (5.3)	487 (5.3)
ERX202485	5902290	1078 (5.3)	1122 (5.3)	1147 (5.3)
ERX202486	2671055	0 (0)	0 (0)	1276 (14)
Control sample				
ERX202487	3526904	0 (0)	0 (0)	0 (0)



Performances:

	k = 1	k = 2
Sensitivity (true positives / total positives)	2/3	3/3
Specificity (true negatives / total negatives)	1/1	1/1
CPU time per read (1 core)	129.2s	192.3s
Time per sample (128 cores)	8.4h	12.5h

Using 512 bit encryption keys, with Intel Xeon E5540 2.53GHz cores:

おわりに

- 個人ゲノム利用が社会にもたらすインパクト
- 利用と保護をどうバランスさせるか
 - 差分プライバシーと秘密計算
- プライバシ研究における機械学習の立位置
 - 攻撃手段
 - 漏えい量評価
 - 秘密計算