Machine Learning meets Privacy Research

佐久間淳 筑波CS/JSTさきがけ

機械学習

データ　　発見・予測

セキュリティとプライバシ

(通信の)セキュリティ

セキュリティとプライバシ

(通信の)セキュリティ

(データ)プライバシ

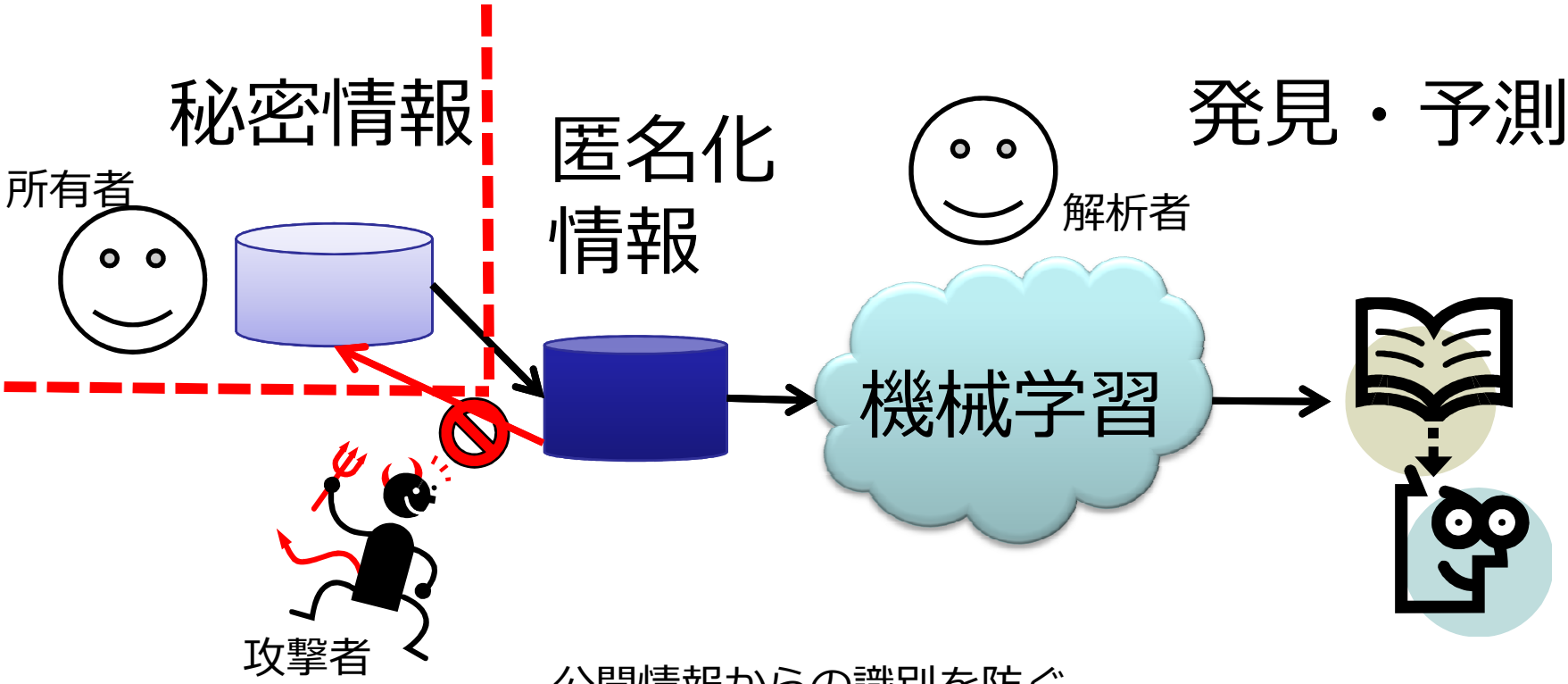
機械学習とセキュリティ・プライバシの接点はどこに？

- **議論としては大きく三つに集約**
  - Input privacy (or 匿名化)
  - Privacy-preserving data mining (or 秘密計算)
  - Output privacy (or 摂動，差分プライバシーなど)
- **クラウド利用(=outsourcing)を考慮すると少し事情は変わる**
  - 詳細は[佐久間11]参照
  - 今回は立ち入らない

Adenda

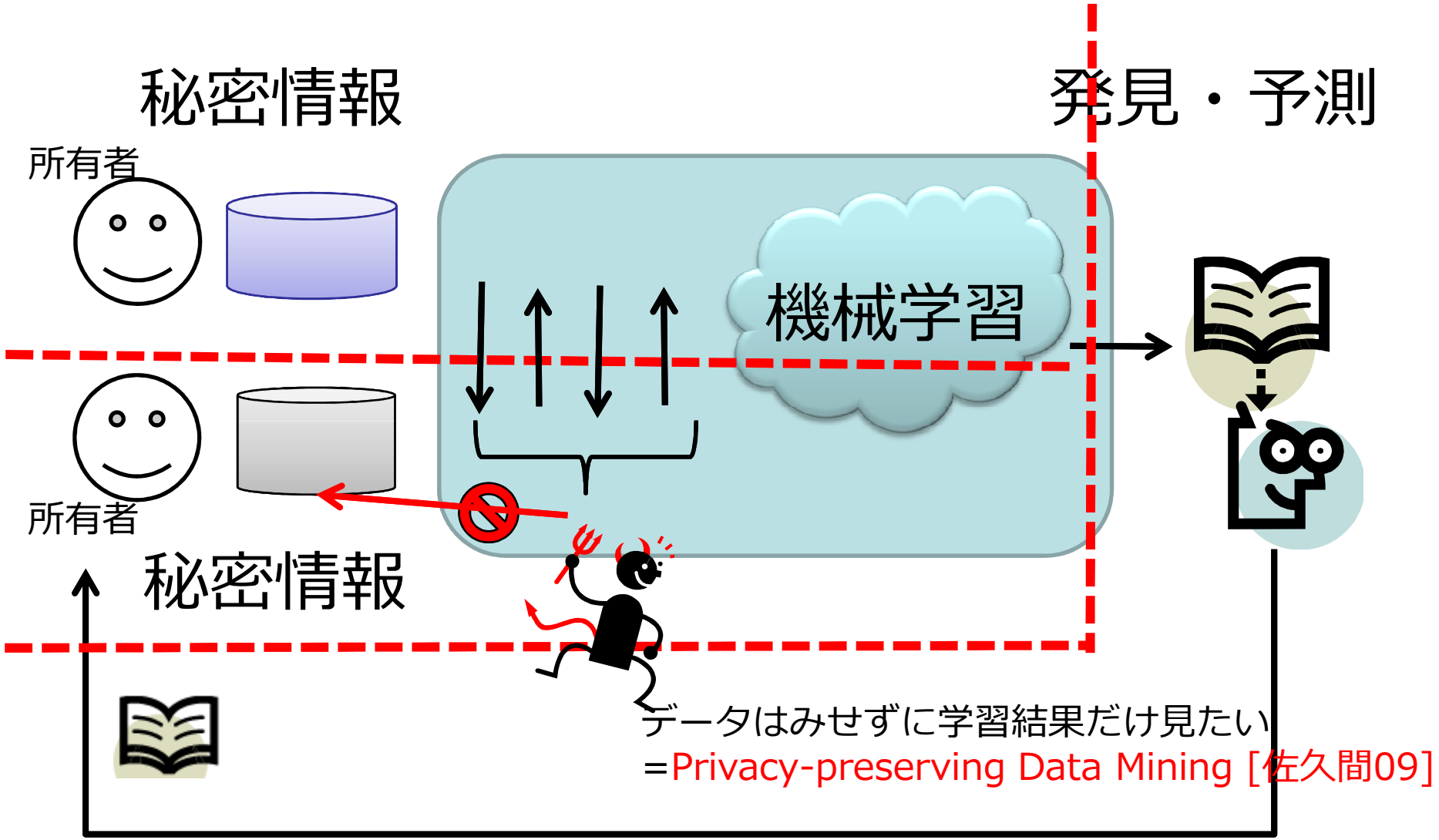
- Privacy researchの三つの方向性
- PPDM：暗号理論アプローチにおけるいくつかのアルゴリズム
- Differential privacyとMLのかかわり
- MLの応用としてのプライバシー研究

Input Privacy

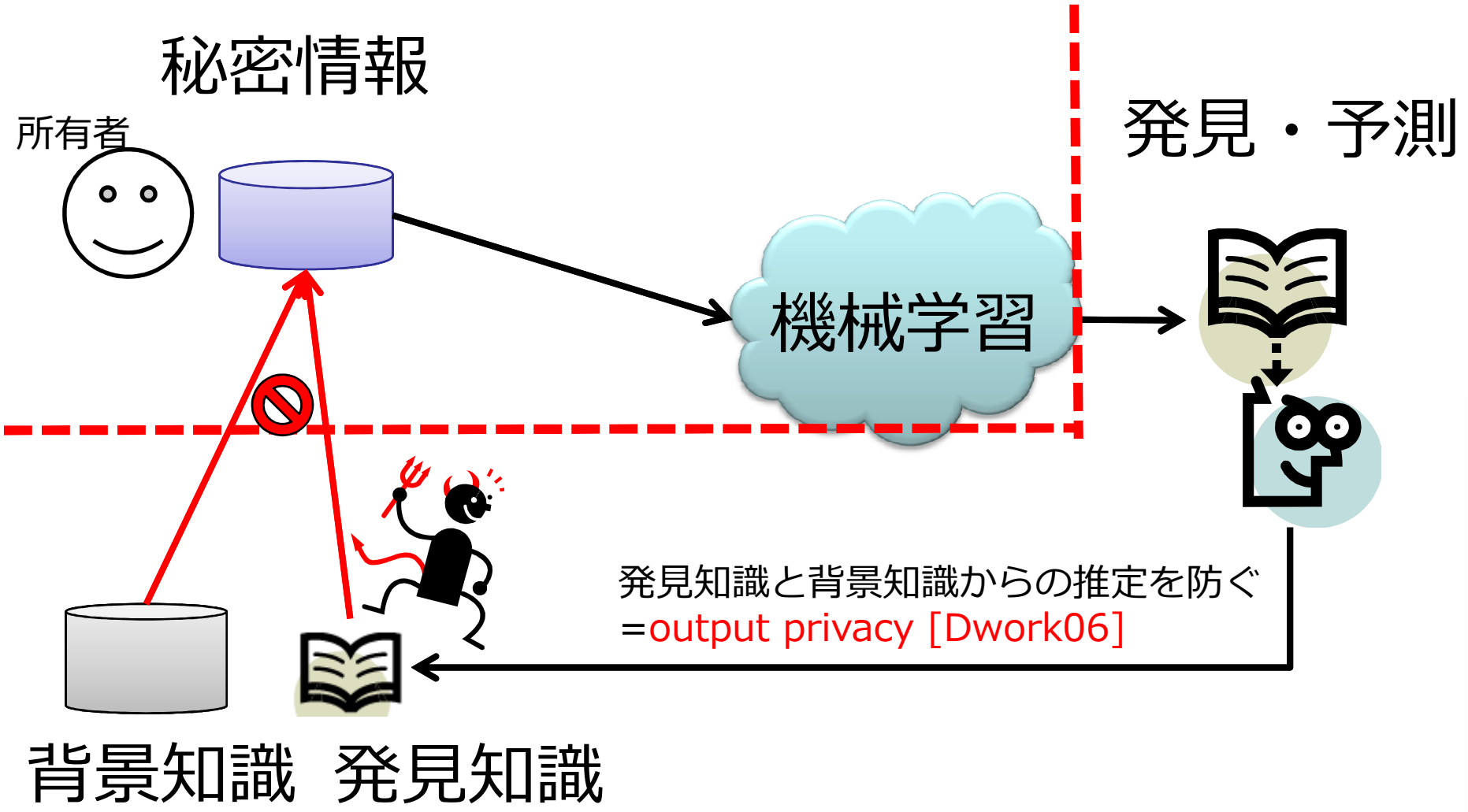


公開情報からの識別を防ぐ
=Anonymization/Input privacy [Sweeney02]

Privacy preserving data mining

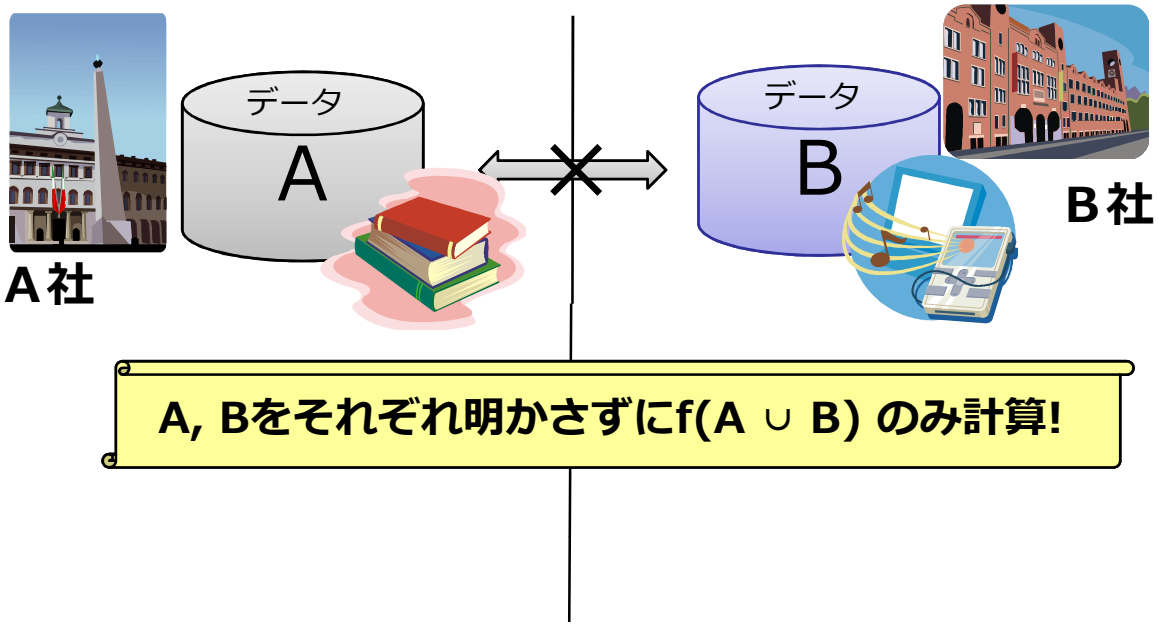


Output privacy



Privacy Preserving Data Mining

Secure function evaluation



- **秘匿関数評価 (Secure function evaluation)**
  - AはデータAをBに開示したくない
  - BもデータBをAに開示したくない
  - ただし結合されたデータA∪Bについてf(A∪B)を知りたい
- **fがデータマイニングの場合，いわゆるPPDM**

準同型性公開鍵暗号

- m ∈ Z_N をメッセージ, r ∈ Z_N を乱数とする
- (pk, sk): 公開鍵と秘密鍵のペア
  - 暗号化: $c \leftarrow \mathrm{Enc}_{pk}(m_0; r_0)$
  - 複号化: $m_0 \leftarrow \mathrm{Dec}_{sk}(c)$
- m0,m1, r1,r2 ∈ Z_N
- 暗号系が(加法的)準同系性を持つとき:
  - 暗号文の和

$$\mathrm{Enc}_{pk}(m_0; r_0) \cdot \mathrm{Enc}_{pk}(m_1; r_1) = \mathrm{Enc}_{pk}(m_0 + m_1; r_1 \cdot r_2)$$

  - 暗号文と平文の積

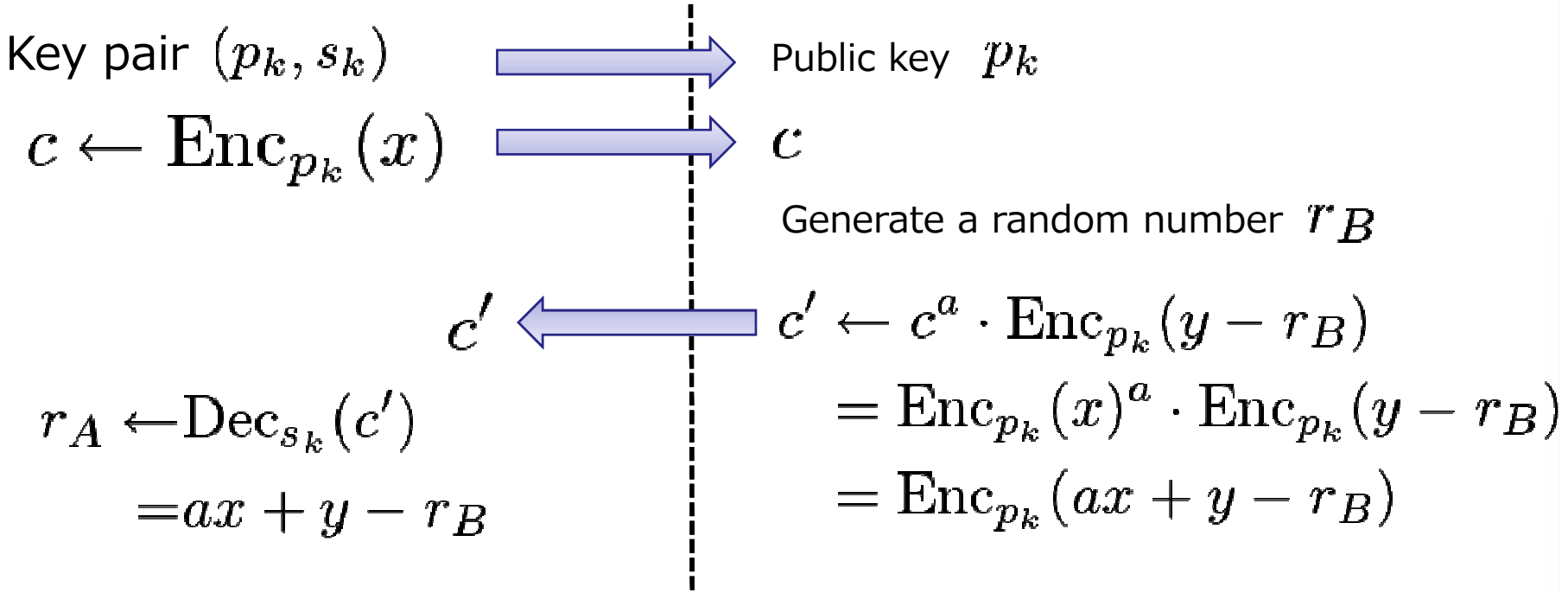
$$\mathrm{Enc}_{pk}(m_0; r_0)^{m_1} = \mathrm{Enc}_{pk}(m_0 m_1; r')$$

e.g. Paillier暗号 [Damgard01]

Example: private computation of ax+y

Alice has x **Bob has y, a**

Problem: compute random shares of ax+y = $r^A + r^B \bmod N$



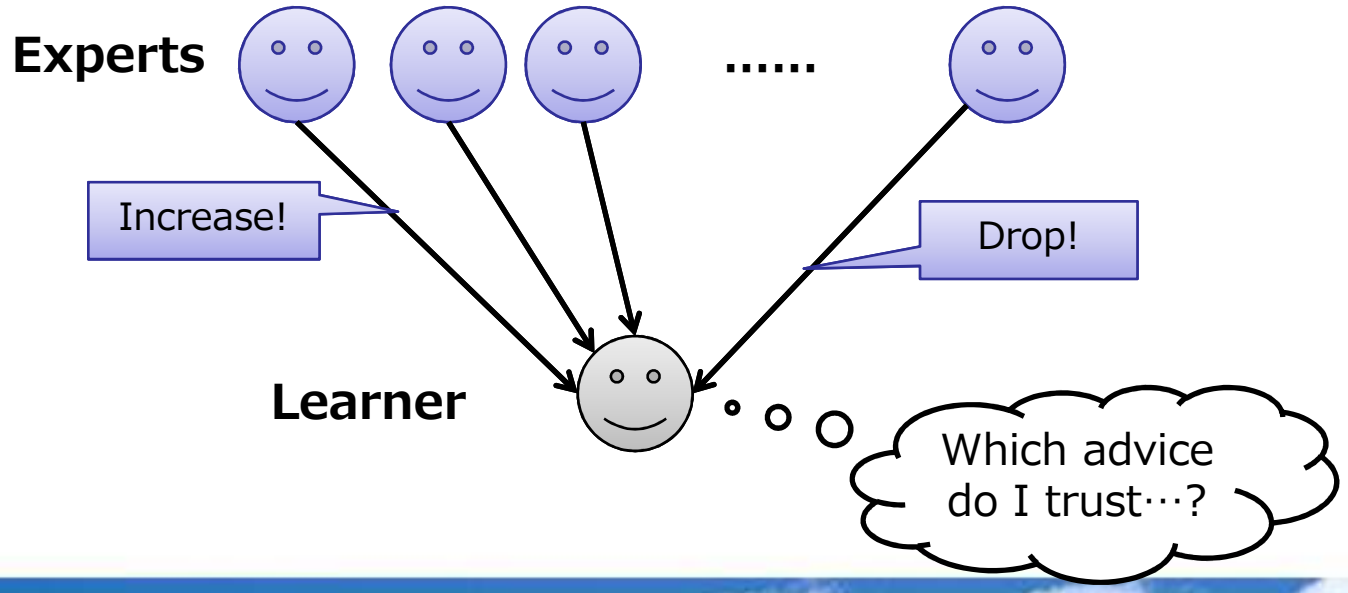
$$r_A + r_B = ax + y \mod N$$

動向

- **準同型暗号を使って相当数のアルゴリズムが”privatize”された**
- **最近はただ準同型暗号上で何かのアルゴリズムをprivatizeするだけではよい会議に通らない**
- **情報観測を制限するモデル自体がターゲットアルゴリズムの発展に新しい知見を与えるなど…**
  - オンライン予測 [Sakuma and Arai, ICML2010]
  - PageRank[Sakuma and Kobayashi, SIGIR2009]
  - ラベル予測[Arai and Sakuma, ECML2011]
  - 結託耐性 [Yang et al., SIGKDD2010]

オンライン予測:株価予測を例に

- **For t=1,2,…,T:**
  1. エキスパート: 忠告「株価は “騰がる ($y_{i,t}$=1)” or “下がる($y_{i,t}$=0)”」を開示
  2. 学習者: エキスパートの忠告を基に予測$y_{H,t}$ を生成
  3. 環境(マーケット): 株価y_t を開示
  4. エキスパート: 自身の予測に対する損失 l (y_t, $y_{i,t}$)を受ける
  5. 学習者: 自身の予測に対する損失 l (y_t, $y_{H,t}$) を受ける



Regret minimization

- **評価基準：regret**
  - 期間Tにおいて、「結果的に最も少ない損失を被ったエキスパート」に比べて、学習者はどの程度損失が多かったか

$$R_{\mathrm{H},T} = \underbrace{L_{\mathrm{H},T}}_{\text{学習者Hの損失和}} - \underbrace{\min_i L_{i,T}}_{\text{最も少ない損失和を被ったエキスパートの損失和}}$$

- **目標:　$R_{H,T} < O(T)$**
  - $T \to \infty$ の極限において$R_{H,T}$ は消失 (a.k.a. **Hannan consistency**)
  - 期間が十分長ければ、最良のエキスパートと同等程度の損失ですむ
- **この目標を達成するために、学習者はどのような戦略をとればよいか？**

Exponential Weighting Scheme

- **戦略**
  - よい予測をしているエキスパートは選ばれやすいように
  - 悪い予測をしているエキスパートは選ばれにくいように

1. **各エキスパートついて重みw_{it}を毎ステップ更新**

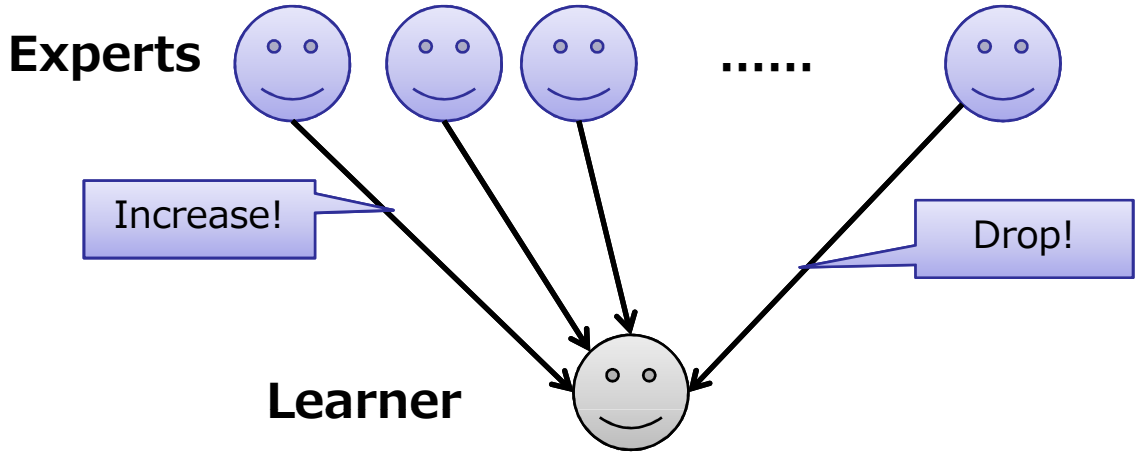
$$w_{i,t} = \exp\left(-\eta \underbrace{\sum_{s=1}^{t-1} \ell(y_{i,s}, y_s)}_{\text{i番目のエキスパートの累積損失}}\right)$$

2. **重みw_{it}を正規化**

$$p_{i,t} = \frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}}$$

3. **p_{it}による予測の決定**
   1. p_{it}に比例する確率でエキスパートを一つ選択 (jとする)
   2. y_{jt}を時刻tの学習者の予測とする

完全情報モデル

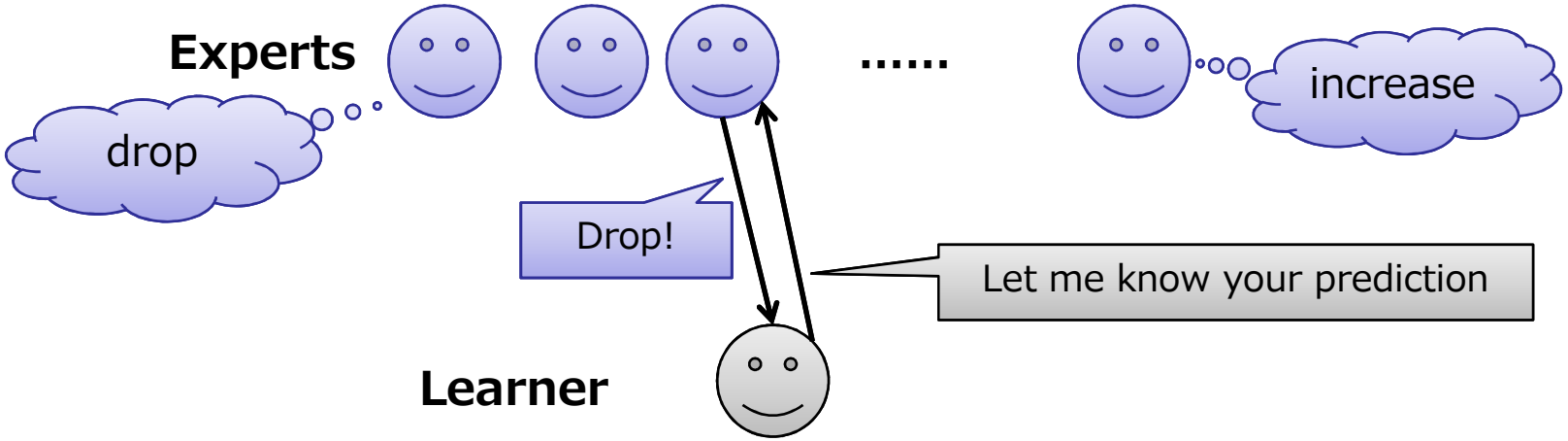


- **完全情報モデル (eg. Exponential weighting)**
  - 学習者はすべてのエキスパートの忠告と損失を観測可能
  - Exponential weighting LearnerのRegret bound[Vovk90]

$$R_{\mathrm{EW},T} \leq \sqrt{2T \ln N}$$ Hannan cosistent!

T:ラウンド数、N: エキスパート数

部分情報モデル



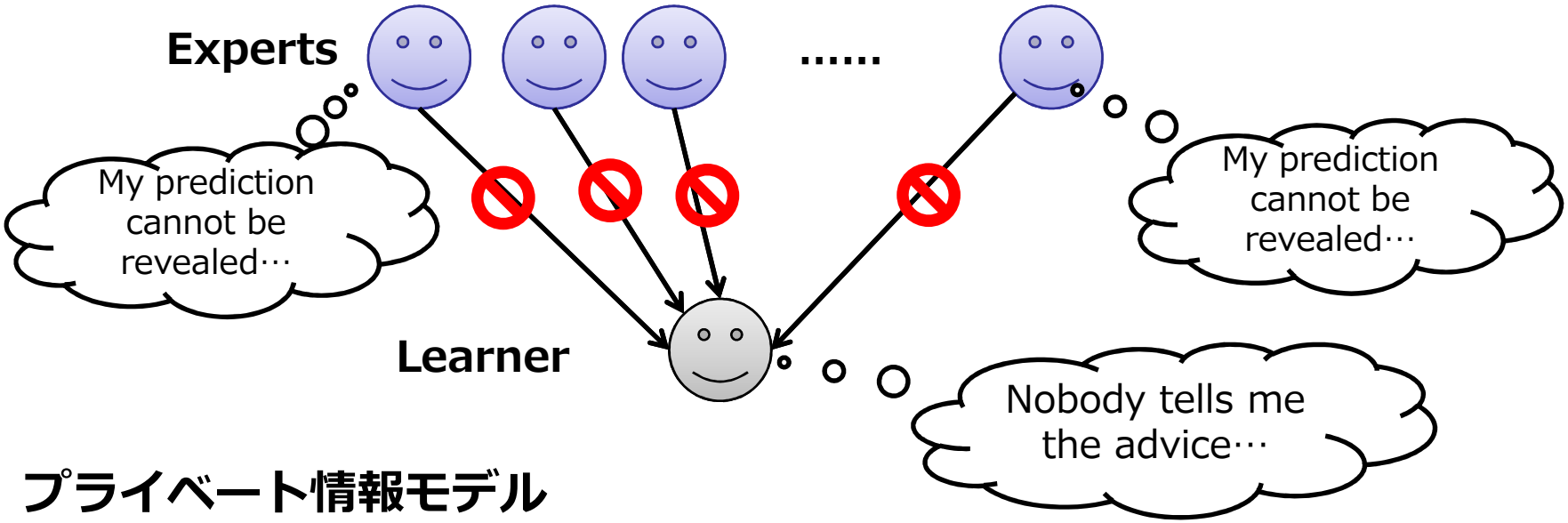
- **部分情報モデル (eg. Exp3 [Auer et. al. 2003])**
  - あらかじめ決めたエキスパートからのみ忠告と損失を観測
  - Exp3 learnerのRegret bound

Hannan cosistent!

$$R_{\mathrm{Exp3},T} \leq 2\sqrt{e-1}\sqrt{NT\ln N}$$

- **エキスパートからの秘密の忠告を扱うにはまだ不足**
- **実現したいシナリオはもっと制限の厳しい情報モデル**

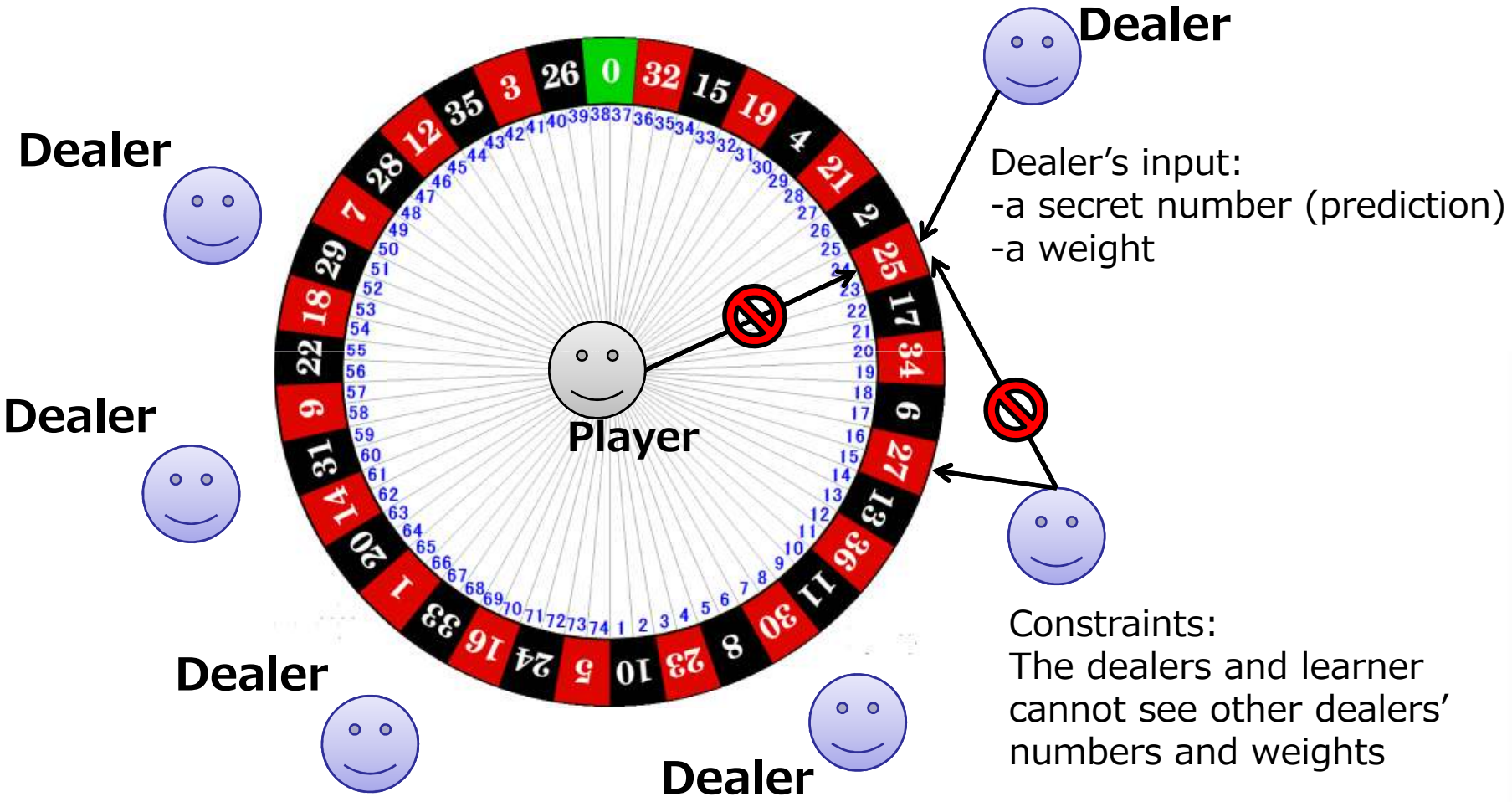
プライベート情報モデル



- **プライベート情報モデル**
  - エキスパートも学習者も互いに予測と損失を一切開示したくない
  - Hannnan consistentなオンライン予測はほとんど不可能に見えるが？

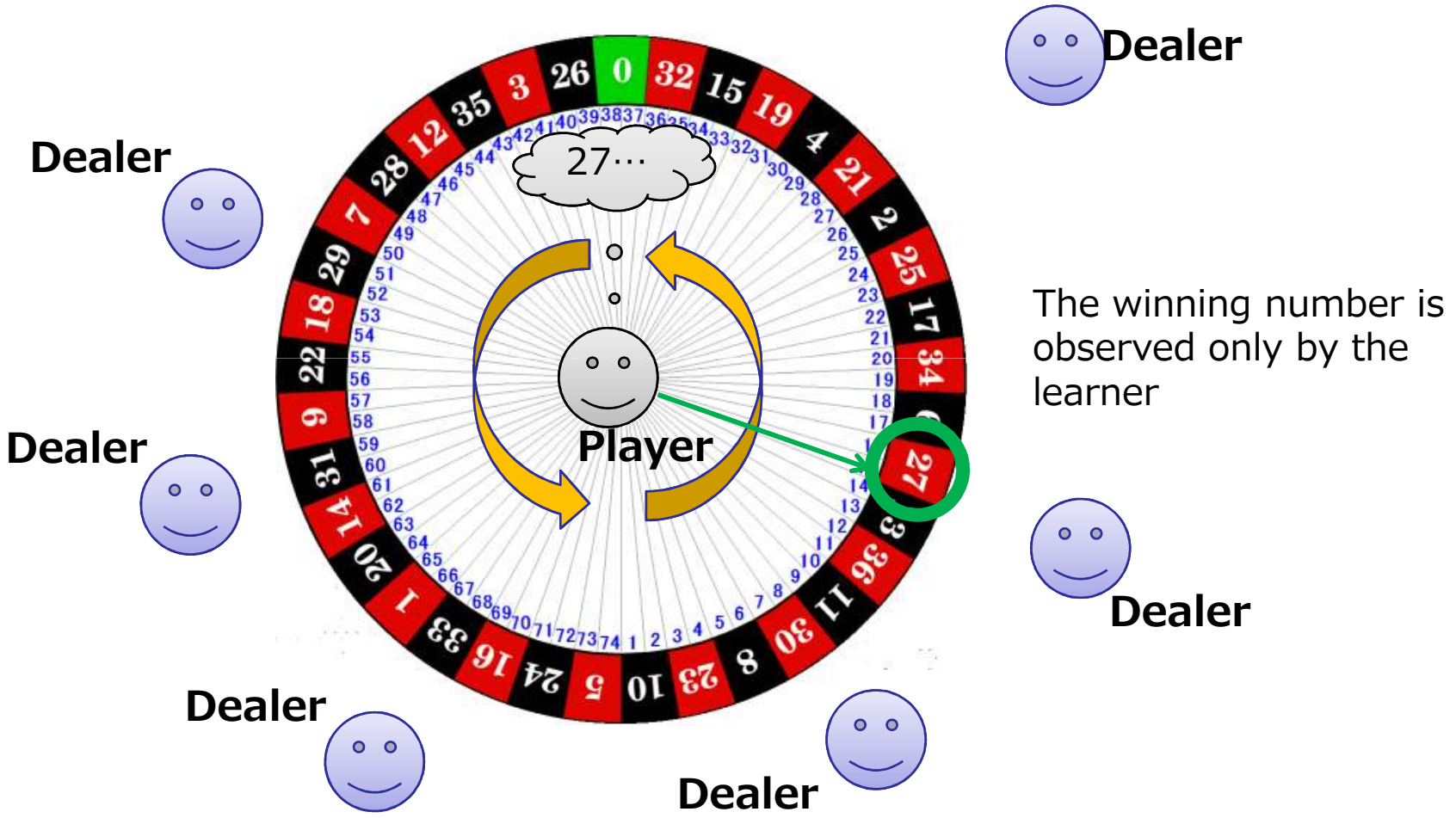
Oblivious Roulette

- **Playing roulette game without seeing the roulette wheel**



Oblivious Roulette

- **Playing roulette game without seeing the roulette wheel**



プロトコルの性質

- **プロトコルの肝**

$$c_{i,k} \leftarrow \left(\mathrm{Enc}_{\mathrm{pk}}(a_{i,k}) \cdot \mathrm{Enc}_{\mathrm{pk}}(-j_k)\right)^{r_{i,k}} \cdot \mathrm{Enc}_{\mathrm{pk}}(y_i)$$

$$= \mathrm{Enc}_{\mathrm{pk}}((a_{i,k} - j_k)r_{i,k} + y_i \mod N)$$

エキスパートの指名jkと学習者の指名が一致すれば0、そうでなければランダムな値をとる

エキスパートiの予測

- **Oblivious roulette protocol**
  - エキスパートiの予測が学習者に届く確率はp_i
  - それ以外のときは乱数のみが届くため、エキスパートは互いの予測や重みを他人に見られない
  - 同様に学習者はだれから予測をもらったか、どんな予測を採用したかエキスパートに知られない
  - これをつかうとexponential weightingをプライベート情報モデルで実行できる

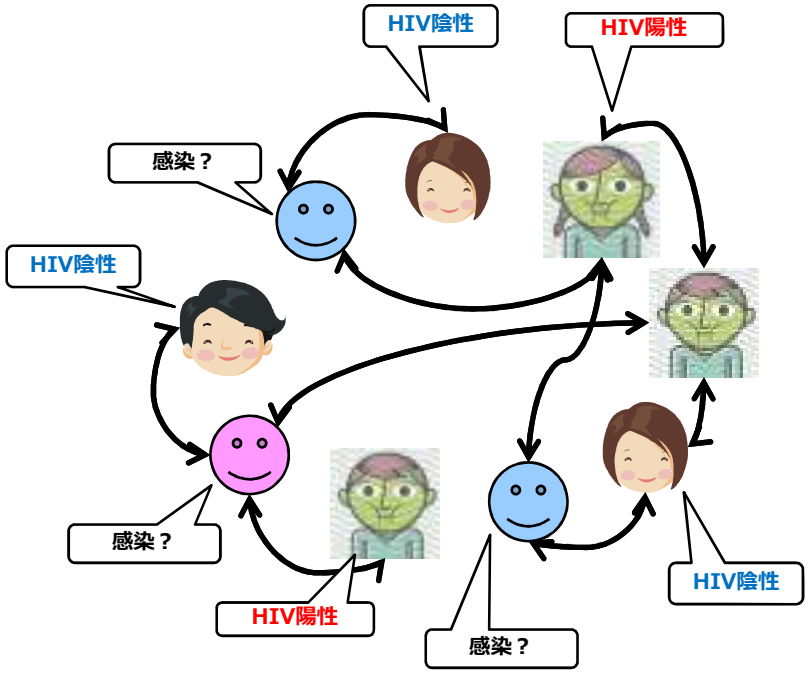
この研究の成果

情報モデル	観測可能な情報	Regret bound
完全情報モデル	全エキスパートの損失,予測	$R_{\mathrm{EW},T} \leq \sqrt{2T \ln N}$
部分情報モデル	指名したエキスパートの損失,予測	$R_{\mathrm{Exp3},T} \leq 2\sqrt{e-1}\sqrt{NT \ln N}$
プライベート情報モデル	**他エキスパートの損失,予測は観測できない**	$R_{\mathrm{EW},T} \leq \sqrt{2T \ln N}$

- プライベート情報モデルにおいても…
  - Hannan consistencyを達成できました
  - しかもregret boundは完全情報モデルと同じオーダーです
- 結論：情報を他のエキスパートと共有しなくとも、学習者は完全情報モデルにおけるexponential weightingと同等の予測ができます

プライバシ保護半教師つき学習(ECML2011, /w H. Arai)

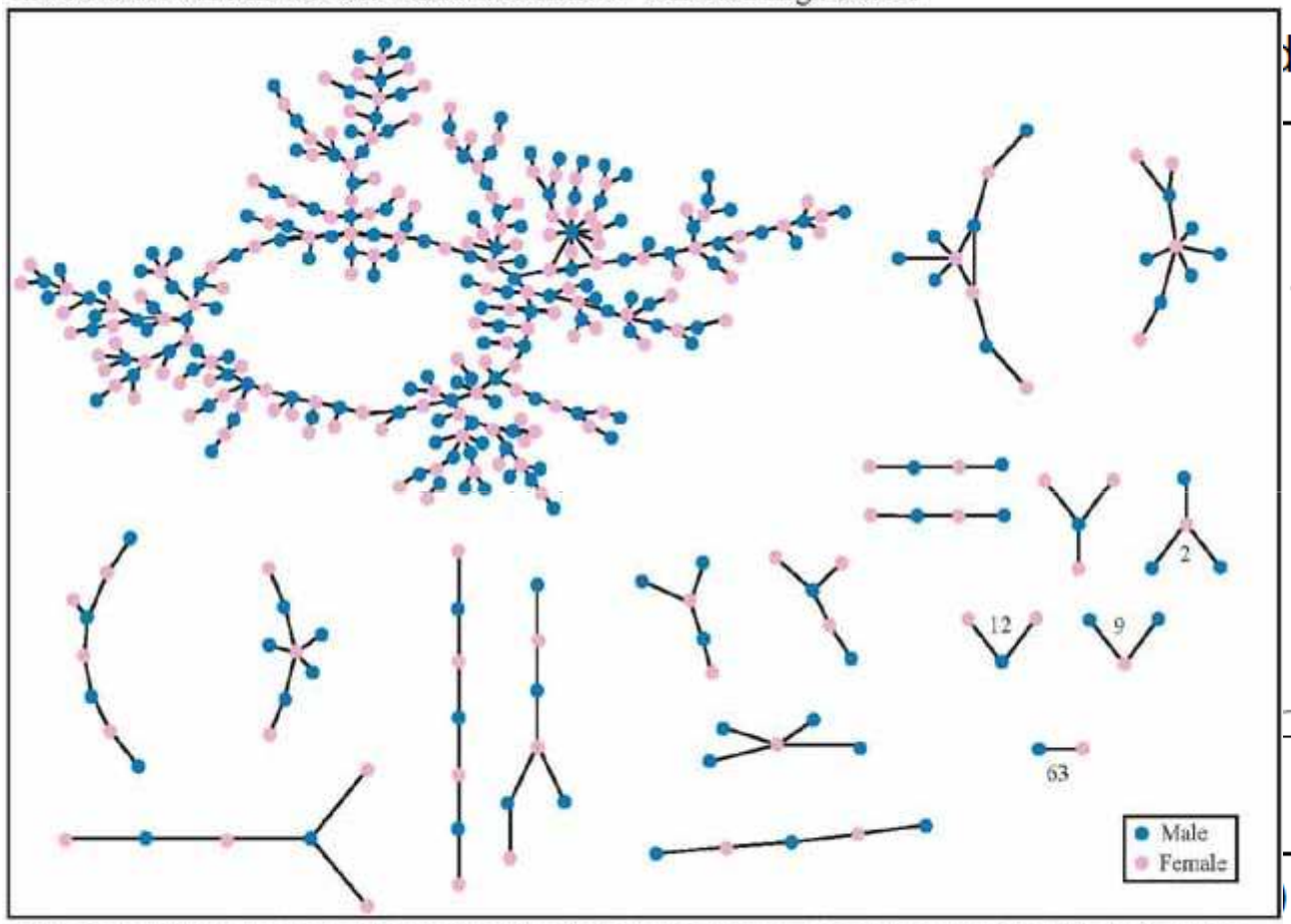
- **伝染病予測のプライバシ**
  - HIVに感染しているか知りたい，でも検査にいきたくない
  - ＡさんはHIVに感染しているかどうか教えてくれない
  - BさんとCさんは過去に性交渉をもったことがあるか教えてくれない
- **秘密の入力**
  - ラベル： Ａさんは HIV陽性か？
  - リンク： BさんとCさんは過去に性交渉をもったか？
- **出力**
  - 私はHIV陽性の可能性があるか？



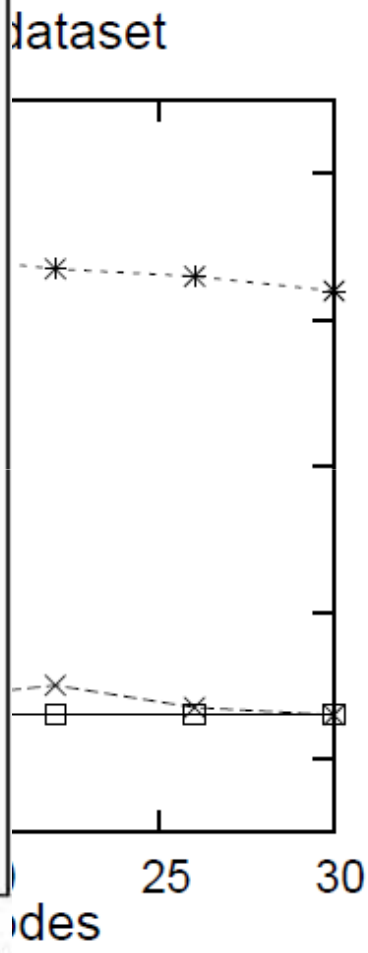
性的交渉ネットワーク

プライバシ保護半教師つき学習(ECML2011, /w H. Arai)

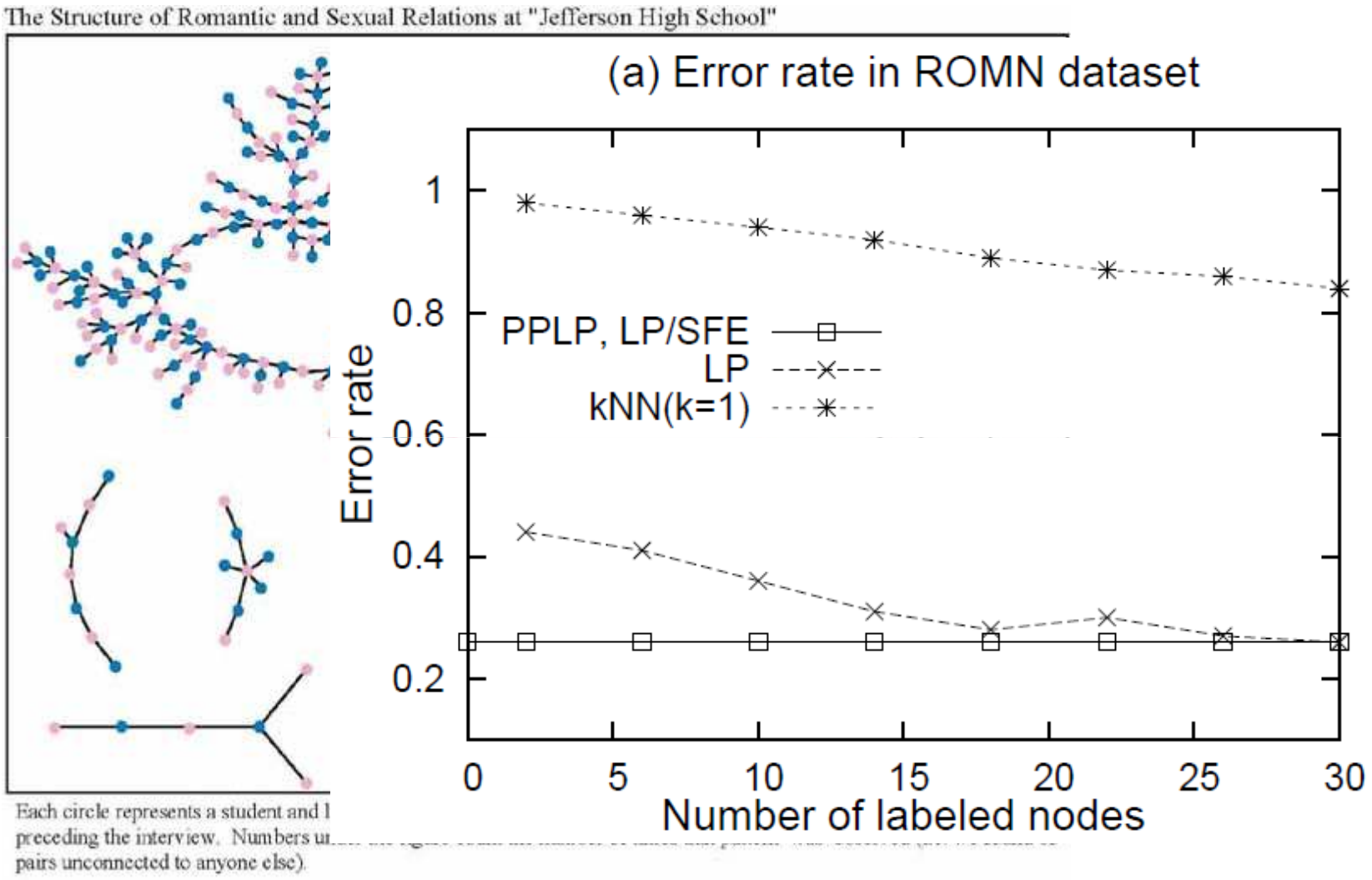
The Structure of Romantic and Sexual Relations at "Jefferson High School"



Each circle represents a student and lines connecting students represent romantic relations occuring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).



プライバシ保護半教師つき学習(ECML2011, /w H. Arai)

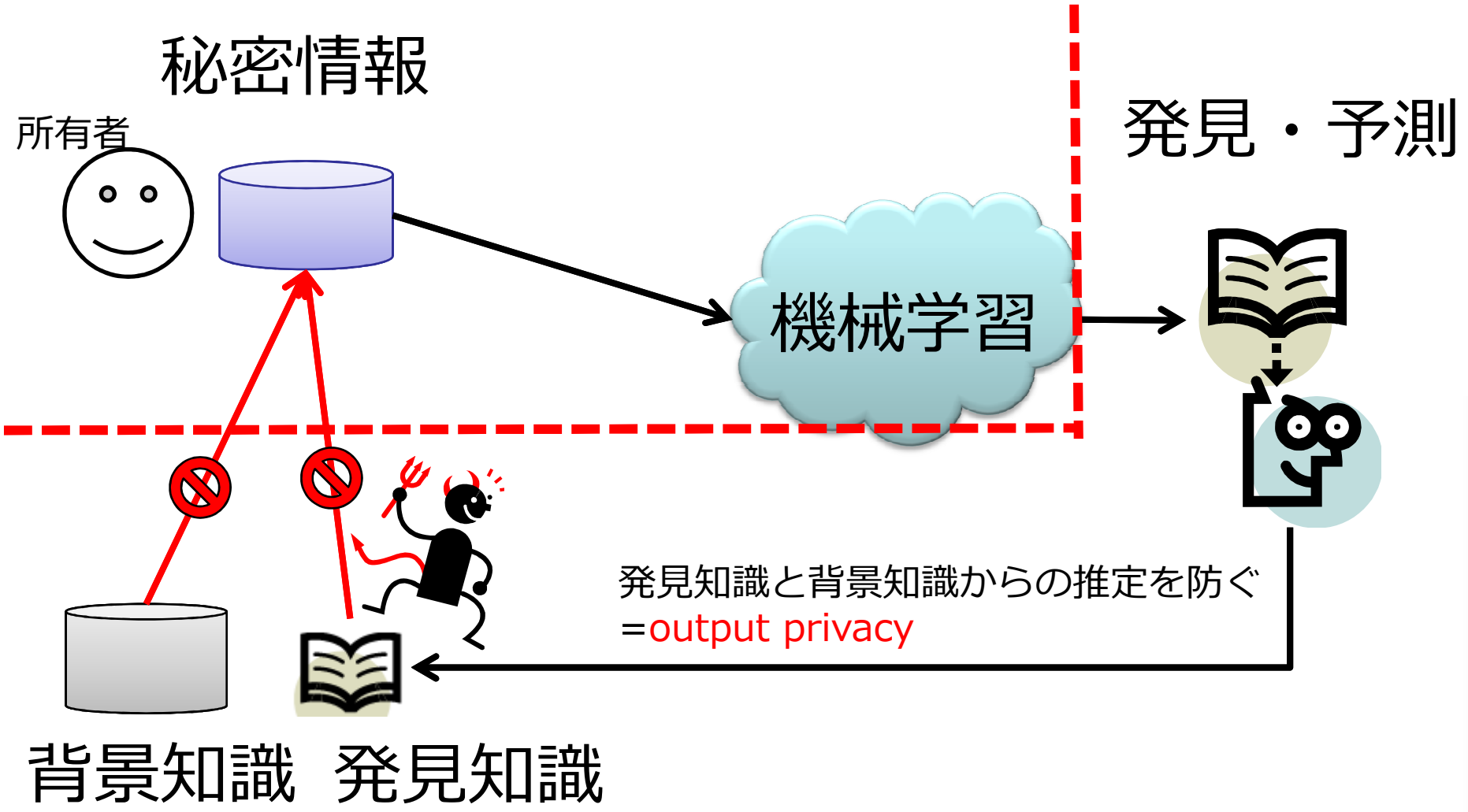


今後の展開

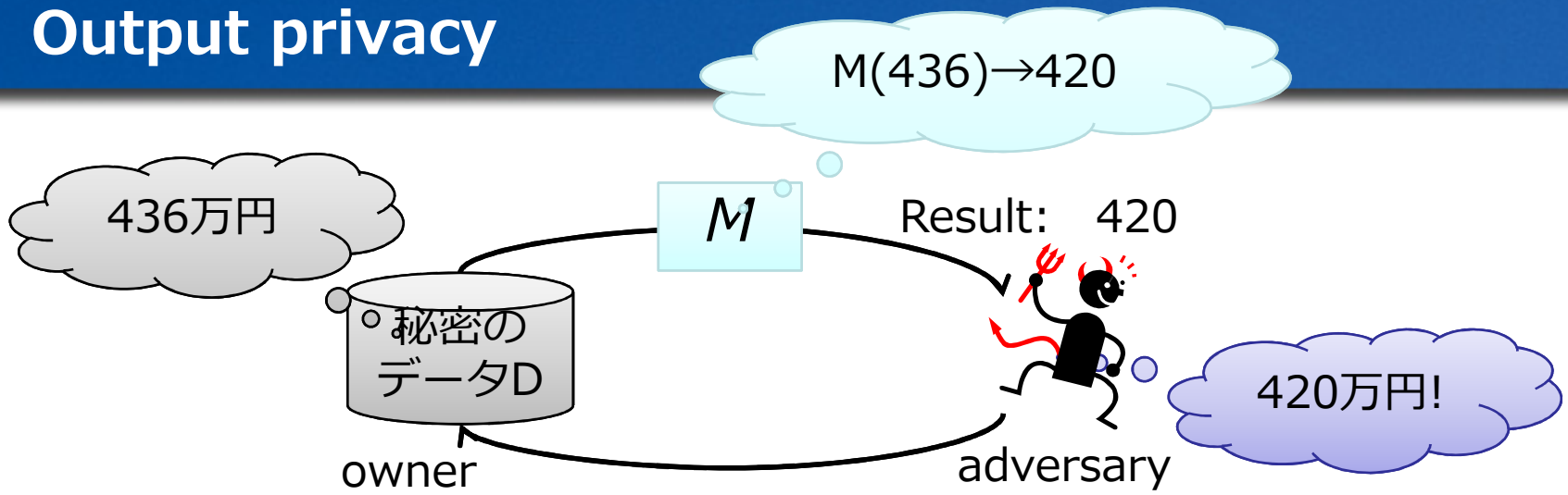
- **利害関係のある複数の主体が出てくるアルゴリズムならなんでも**
  - リンク予測
  - 安定結婚
  - 動的計画法
  - クラウドソーシング
  - 予測市場
  - オンライン学習,ゲーム(均衡解探索)
  - 詳しい方いっしょにやりませんか？
- **実際にうごかすプラットフォームが必要**
  - Androidで実装計画中

出力プライバシー

Output privacy



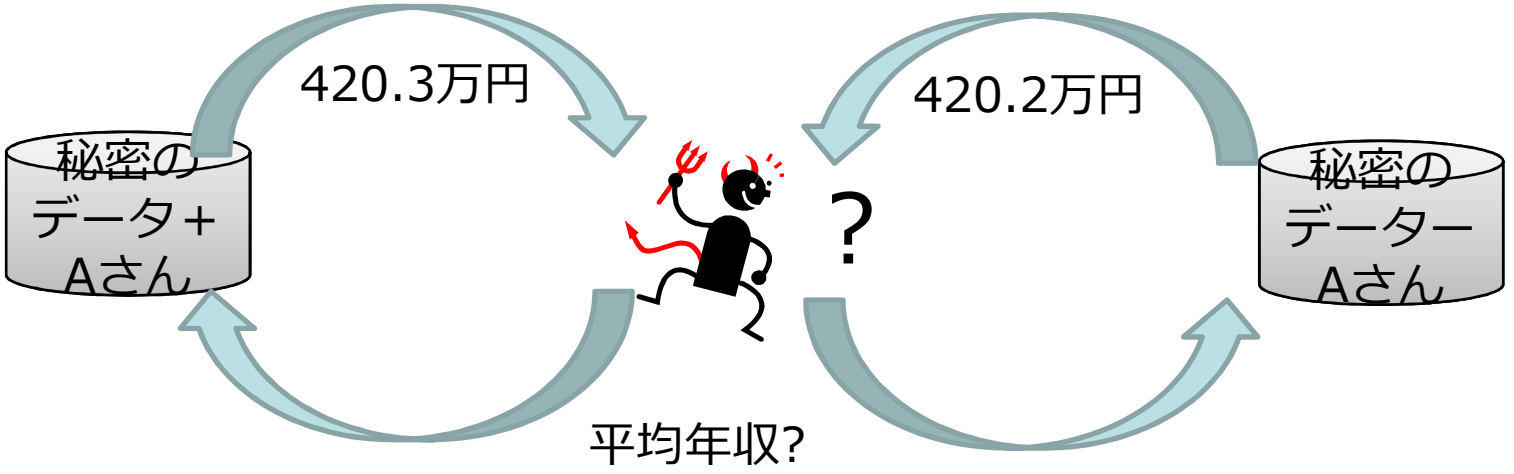
Output privacy

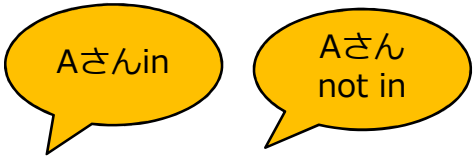


Query:入社3年目の平均年収?

- **攻撃者はクエリを発行**
- **オーナーはf(D)を計算**
- **Dの秘密を守るためにf(D)に細工Mをしてから開示**
- **攻撃者はM(f(D))からDを推測**
- **問題：どんなMなら安全なのか？**

Output privacyにおける Semantic security





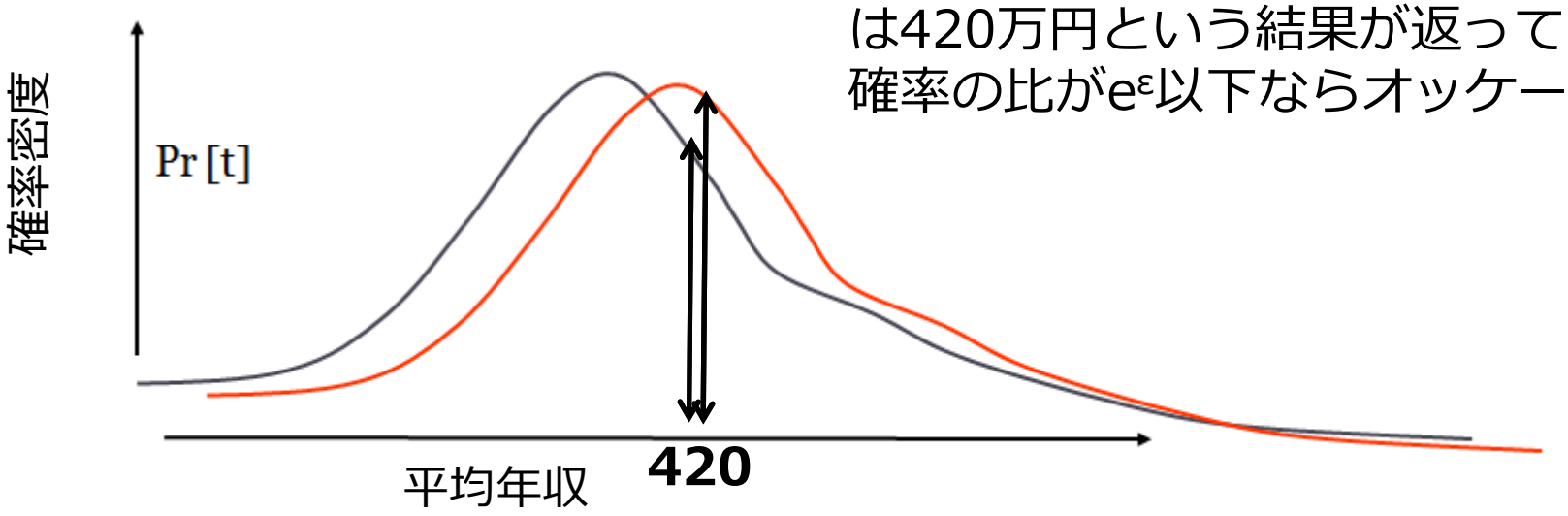
- f(D+A)とf(D)がそんなに変わらなければ、f(D+A)の開示はAさんのプライバシを侵害していない (ことにしよう！)

Differential Privacy[Dwork06]

- そんなに変わらないとは…?
- Differential Privacy[Dwork06]

$$\frac{\Pr[\mathcal{K}(DB - You) \in S]}{\Pr[\mathcal{K}(DB + You) \in S]} \leq e^{\varepsilon}$$

Aさんがいるといないとで、平均年収は420万円という結果が返ってくる確率の比がe^{ε}以下ならオッケー

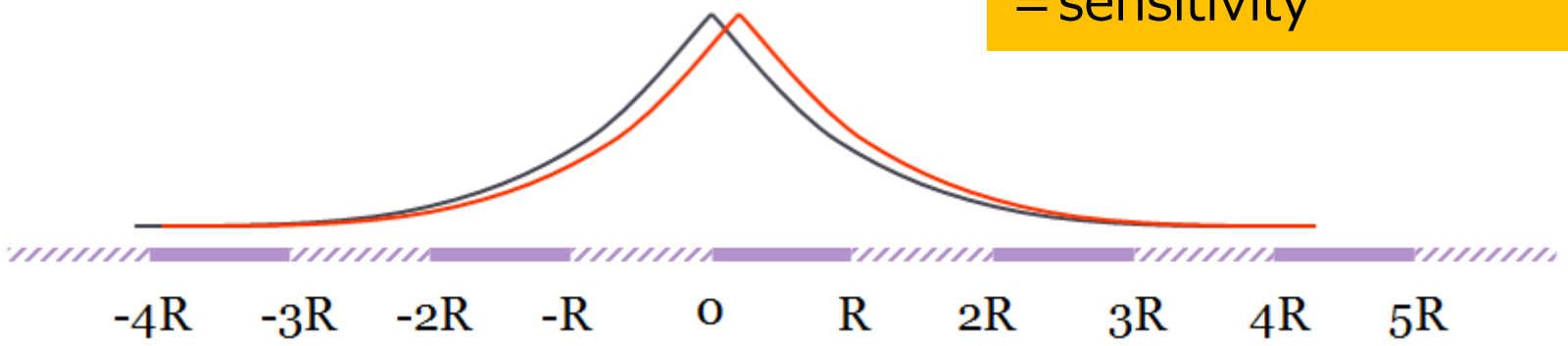


どうやってDPを実現するか? Laplaceメカニズム

- f(DB)に加法的ノイズを加えて開示
- 定理： R = Δf/εとしたときにLap(|x|/R)に従うノイズrを加えればε-differential privacy
- Δf：fのデータの変化に対する感度

$$\mathrm{Lap}(x;R) = \frac{1}{2R} e^{-\frac{|x|}{R}}$$

Aさんがいるといないとで
結果がどれくらい変わるか、
で決まる定数
＝sensitivity



Sensitivity

- **ある人がいるかいないかで、計算結果は最大でどれだけかわるか？**

$$\Delta f = \max_{DB,\ You} |f(DB+You) - f(DB-You)|$$

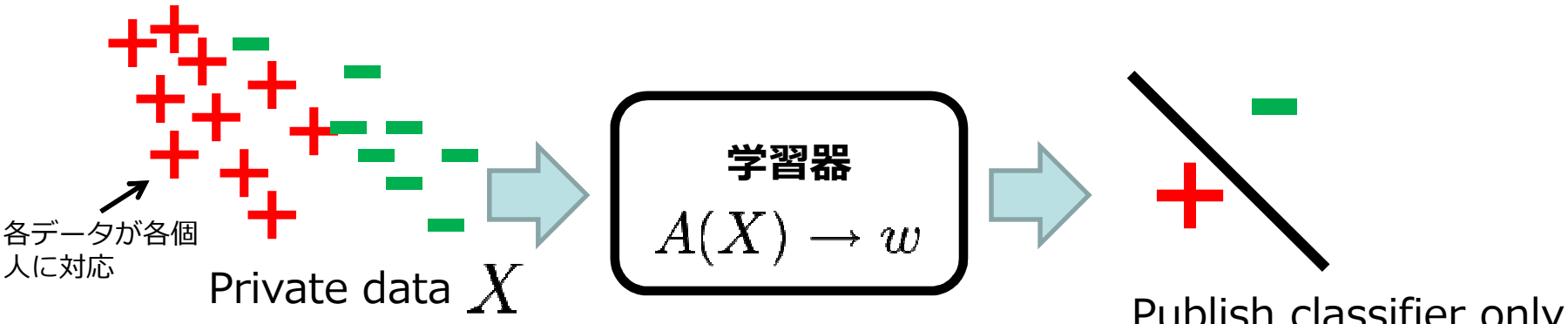
- **例**
  - Countクエリ：　Δf=1
  - 平均クエリ：Δf=S/N, S:値のレンジ、N:人数
  - maxクエリ：Δf=S, S:値のレンジ

複雑さの抑制とプライバシ保護

- **機械学習**
  - ロス最小化したい
  - 学習結果の複雑さを抑制したい：正則化, 低ランク近似, etc
- **プライバシ保護**
  - 出力の秘密情報への適合を抑制したい：匿名化，差分プライバシ
  - データが持っている効用最大化したい
- **両者は結構似ている？**
  - 「複雑さの抑制」と「出力の秘密情報への適合の抑制」は同時に達成できるのでは？
  - 機械学習のテクニック(e.g.正則化, 低ランク近似)がプライバシ保護ツールとして有用なのでは？

Differentially private ERM minimization [Chaudhuri11]@JMLR

- **Private dataからの分類器学習**



- **Differential privacy in ERM**
  - 分類器 w : $f(x) = w^T x$
  - あるデータがあろうとなかろうと, w が学習される確率はあまり変わらない

$$\frac{Pr[A(X) = w]}{Pr[A(X') = w]} \leq \exp \epsilon$$

Differentially private ERM minimization [Chaudhuri11]@JMLR

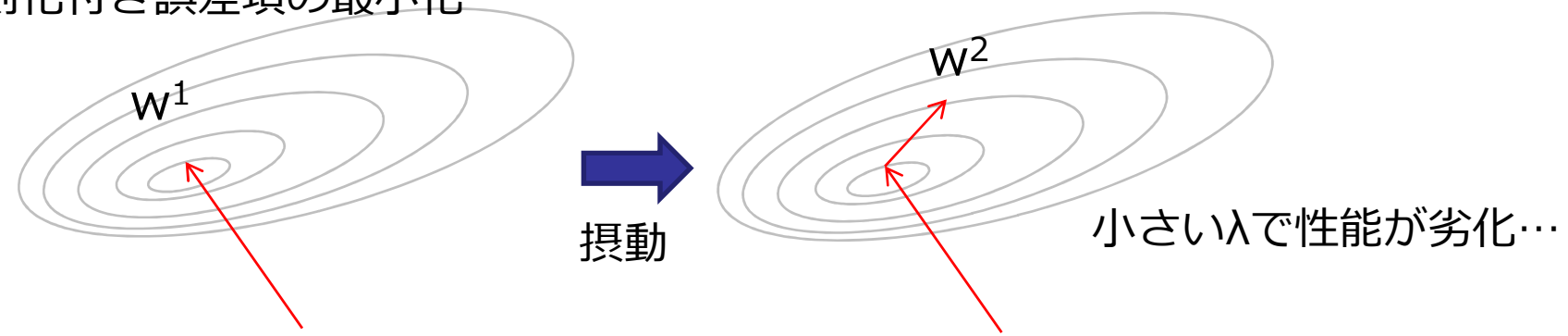
- データD={$(x_1,y_1),\cdots,(x_N,y_N)$}, $||x_i||<1$, $y_i \in \{-1,1\}$
- w^1: Dからlogistic regressionによって求めた分類器

$$\hat{L}(w) = \frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i w^T x_i})$$

$$\hat{f}_\lambda(w) = \frac{1}{2}\lambda||w||^2 + \hat{L}(w)$$

- w^1の代わりに、differential privateな分類器w^2を開示したい
  - $\Delta f = n\lambda/2$, $R = \Delta f/\varepsilon$, $r \sim Lap(w; R)$として$w^2=w^1+r$を開示

正則化付き誤差項の最小化



Differentially private ERM minimization [Chaudhuri11]@JMLR

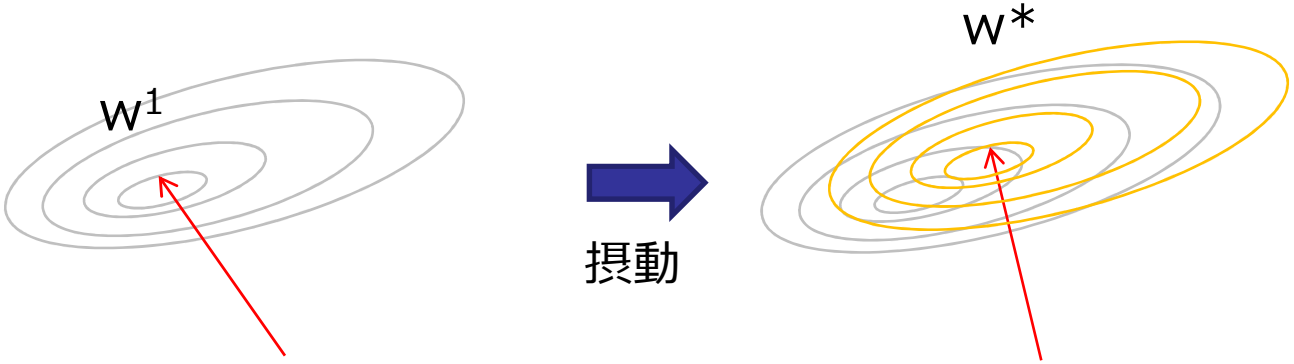
- アイディア: ノイズを加えた目的関数を最適化
- ベクターb ～ $h(b) \propto e^{-\frac{\epsilon}{2}||b||}$
- 以下の問題を解く

$$w^* = \mathrm{argmin}_w \frac{1}{2}\lambda w^T w + \frac{b^T w}{n} + \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i}).$$

Cf. $\hat{f}_\lambda(w) = \frac{1}{2}\lambda||w||^2 + \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i})$

- Theorem2.

  w*は ε-differentially private



Differentially private ERM minimization [Chaudhuri11]@JMLR

- **本質的な解釈**
  - ノイズをのせる→解をぼやけさせる
  - 正則化項を加える→解をなまらせる
  - どうせ似たような事をやっているのならまとめてやって精度の高いものを
- **ERMにおいてε-differential privacyを与える問題が簡単に定義できるというのはなかなか強い結果**

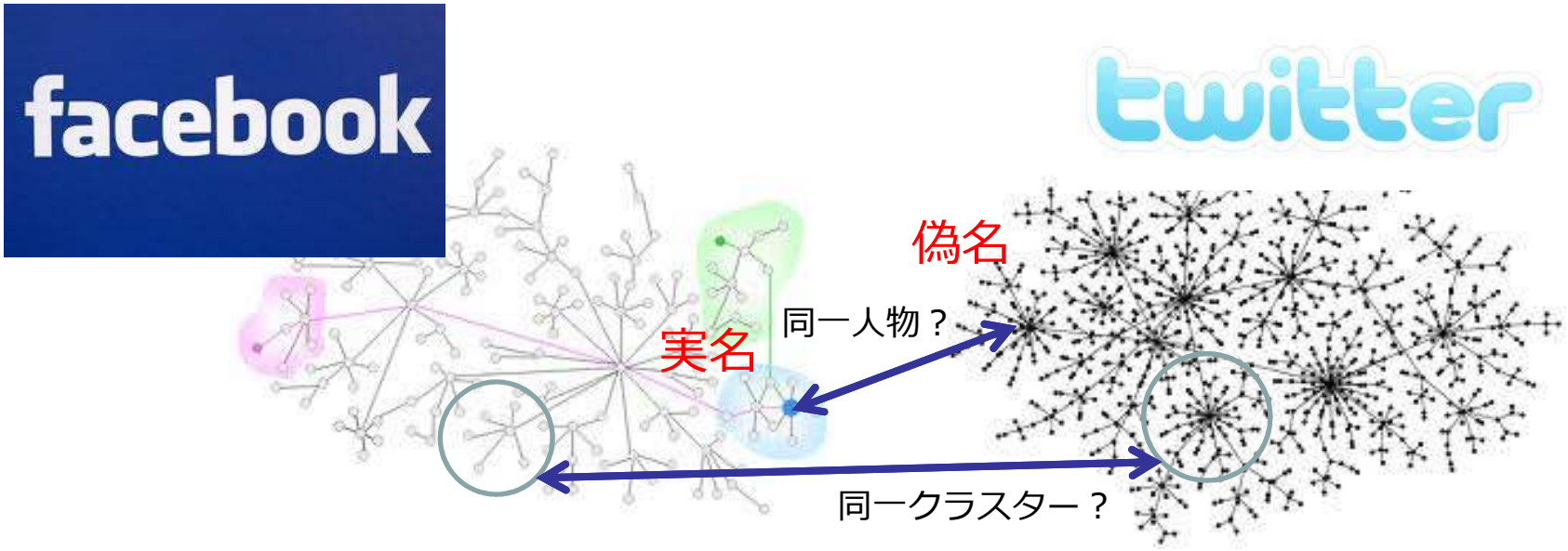
MLのprivacy researchへの応用可能性

AdversaryとadministratorのAttack

	被害	攻撃ツール	防御ツール
暗号	盗聴	数論	数論
ウイルス /malware	感染	システム脆弱性解析	保護ソフトウェア
ネットワーク セキュリティ	侵入	ネットワーク脆弱性解析	保護ソフト/ハードウェア
SNS/web サービス	脱匿名化, 属性推定	クローリング＆ 機械学習?	機械学習？

- **侵入検知や異常検知，スパムフィルタが機械学習の重要なアプリケーションであったように…**
- **脱匿名化や属性推定, anti-脱匿名化やanti-属性推定が機械学習の主要なアプリになるのでは？**

SNS脱匿名化とanti脱匿名化



- facebookでは顕名，twitterでは偽名
- Twitterには実名ノードも多数あり
- 周辺ノードのリンク情報はvisible
- 攻撃側の問題: あるfacebookアカウントに対応するtwitterアカウントを同定できるか？
- 防御側の問題：自分のfacebookアカウントが同定されないためにはどのような属性やリンクを作ればよいか？

推薦のトランザクション推論とantiトランザクション推論

- **“You might also like:” Privacy Risks of Collaborative Filtering，IEEE S&P2010 [Calandrino2011]**
  - Webなどから推薦情報やセールスランキングなどののダイナミクスを収集し，それを目標ユーザの”外部情報” と結び付け，目標ユーザの未知のトランザクションを推論
- **解析例：**Ms. BrownはR&Bのアルバムにa1,a2,a3を含むレビューを書いている．アイテムtは古いR&Bのアルバムである．Day1では70,000位だったが，Day2には15,000位に上昇．Day2では，アイテムtはa1,a2の類似アイテムリストに出現し，a3の類似アイテムリスト内にて類似度が上昇した．これらからMs. Brownはアイテムtを購入したと推測できる．実際，1月以内にMs. Brownはアイテムtのレビューを書いた
- **非常に素朴な推論を使っている**
- **MLを使えばもっと高精度な同定が可能？**
- **トランザクション推論を妨げるにはどんな推薦が有効か？**

おわりに

- MLとprivacy researchには多様なかかわり方がある
- 理論：抽象化・単純化とプライバシ保護の根源的な同一性
- アルゴリズム：privatizationがあたらしい設定のMLを作る
- 応用：SNS/webサービスのanti-XXはpromisingなアプリ？

関連文献

- [佐久間11] 佐久間淳,高橋克巳,クラウドストレージにおける個人情報の利活用とプライバシー保護,情報処理, Vol.52 No.6, 2011
- [Sweeney02] Sweeney, L. : k-Anonymity : A Model for Protecting Privacy,
- World , Vol.10, No.5, pp.557-570（2002）.
- [佐久間09] 佐久間淳，小林重信：プライバシー保護データマイニング，人
- 工知能学会誌，Vol.24, No.2, pp.283-294（2009）.
- [Dwork06] Dwork, C., McSherry, F., Nissim, K. and Smith, A. :
- Calibrating Noise to Sensitivity in Private Data Analysis,
- Theory of Cryptography , pp.265-284（2006）.
- [Damgard01] Damgard, I. and Jurik, M. A Generalisation, a Simplication and Some Applications of Paillier's Probabilistic Public-Key System. In Public Key Cryptography, pp. 119–136. Springer, 2001.
- [Sakuma10] Jun Sakuma, Hiromi Arai: Online Prediction with Privacy. ICML 2010: 935-942

- [Sakuma09] Jun Sakuma, Shigenobu Kobayashi: Link analysis for private weighted graphs. SIGIR 2009: 235-242
- [Arai11] Hiromi Arai, Jun Sakuma: Privacy Preserving Semi-Supervised Learning for Labeled Graphs, ECML2011, to appear.
- [Yang10] Bin Yang, Hiroshi Nakagawa, Issei Sato, Jun Sakuma: Collusion-resistant privacy-preserving data mining. KDD 2010: 483-492
- [Chaudhuri11] Kamalika Chaudhuri, Claire Monteleoni, and Anand Sarwate. Differentially Private ERM , vol. 12, pp. 1069−1109, JMLR, 2011.
- [Calandrino2011] Joseph A. Calandrino, Ann Kilzer, Arvind Narayanan, Edward W. Felten, and Vitaly Shmatikov"You Might Also Like:"Privacy Risks of Collaborative Filtering, IEEE S&P 2011, to appear.