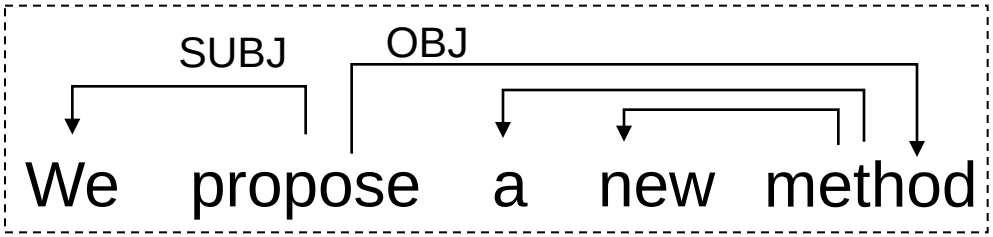
I-1

模倣学習による依存構造解析

坪井 祐太（日本アイ・ビー・エム株式会社）

- 依存構造解析(係り受け解析)とは？
  - 文中の単語間の修飾関係を表す木構造を予測



  - 評判分析・関係抽出・機械翻訳などの基盤技術
- 提案手法の効果
  - 少しの精度低下で、**数十倍の高速化**を実現
    - 英語ベンチマークデータでの評価 (Penn Treebank)

	既存手法 (Zhang&Nivre2011)	提案手法
解析速度	27文/秒	**830文/秒**
精度（UAS）	92.9%	90.7%

  - 応用：マイクロブログ等の大規模テキストデータの処理

I-1

模倣学習による依存構造解析

坪井 祐太（日本アイ・ビー・エム株式会社）

- Transition-based Dependency Parsing（既存手法）
  - 文の前から順に係り受け関係を決定（マルコフ決定過程）
    - 正解データを用いて方策を教師付き学習
  - 解析時の**誤差伝播**が課題
    - 予測履歴も特徴に用いるため、前の誤りが伝播する（Non-i.i.d.）
    - 既存手法では大域最適化&ビーム探索によって回避 → 探索幅に比例して解析速度低下

    →正解データ下の状態分布と学習した方策が観測する状態分布が異なる
- 模倣学習(Imitation-learning)による誤差伝播回避
  - 強化学習問題としての依存構造解析器学習:
    - 膨大な状態空間数（状態を表す特徴次元≒約500万）
    - エピソード単位の決定過程
    - 報酬を最大化する方策（オラクル）が利用可能
  - 模倣学習（DAGGER, Ross et al. 2011）の適用
    - 方策空間の中で報酬の大きい領域のみを探索→学習の効率化
    - **解析速度を落とすことなく精度が向上**（89.7%→90.7%）

I-2

階層ベイズモデルによる消費者異質性を考慮したクロスメディア効果推定

日高徹司（博報堂）、佐藤忠彦（筑波大学）

目的

1. 大規模データでさえ捉えることが困難な，インターネット、テレビ、ラジオ、新聞、雑誌、屋外広告などの複数のメディアの相乗効果（クロスメディア）効果を小規模な実験で測定
2. 各消費者のデモグラフィック特性やブランドの知識や態度（興味、魅力など）と広告効果の異質性、多様性の関係を探る

調査手法

1. Web調査を用いた広告の強制露出実験
2. 被験者数：各グループ100人×16グループ=1600人
3. 各ブランド5つの広告素材（TVCM，新聞広告，雑誌広告，交通広告，Webサイトなど）
4. 露出前後にブランドの知識や態度を質問

	A	B	AB	C	AC	BC	DE	D	AD	BD	CE	CD	BE	AE	E
1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
2	○	○	○	○	○	○	○	×	×	×	×	×	×	×	×
3	○	○	○	×	×	×	×	○	○	○	○	×	×	×	×
4	○	○	○	×	×	×	×	×	×	×	×	○	○	○	○
5	○	×	×	○	○	×	×	○	○	×	×	○	○	×	×
6	○	×	×	○	○	×	×	×	×	○	○	×	×	○	○
7	○	×	×	×	×	○	○	○	○	×	×	×	×	○	○
8	○	×	×	×	×	○	○	×	×	○	○	○	○	×	×
9	×	○	×	○	×	○	×	○	×	○	×	○	×	○	×
10	×	○	×	○	×	○	×	×	○	×	○	×	○	×	○
11	×	○	×	×	○	×	○	○	×	○	×	×	○	×	○
12	×	○	×	×	○	×	○	×	○	×	○	○	×	○	×
13	×	×	○	○	×	×	○	○	×	×	○	○	×	×	○
14	×	×	○	○	×	×	○	×	○	○	×	×	○	○	×
15	×	×	○	×	○	○	×	○	×	×	○	×	○	○	×
16	×	×	○	×	○	○	×	×	○	○	×	○	×	×	○

直交表L16（○：広告呈示，×：呈示せず）
A,B,・・,E：広告素材
AB,AC,・・,AE：交互作用

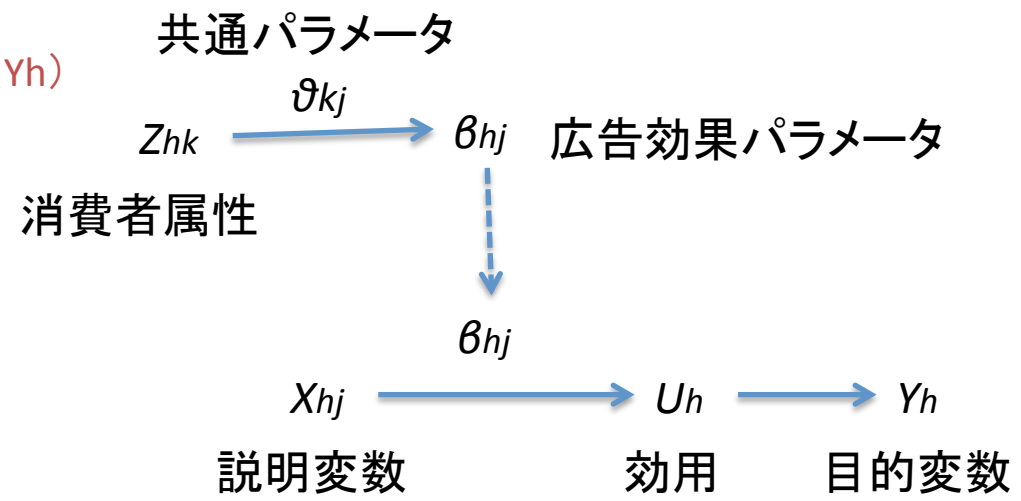
I-2

階層ベイズモデルによる消費者異質性を考慮したクロスメディア効果推定

日高徹司（博報堂）、佐藤忠彦（筑波大学）

モデル

1. 階層ベイズ二項ロジットモデルを適用
   1. 目的変数：広告露出後のブランド態度（Yh）
   2. 説明変数：広告露出の有無（Xhj）
   3. 個人属性：Zhk
2. 異質性を表す添え字
   1. h：消費者
   2. k：個人属性ID
   3. j：広告素材



得られた知見

1. 広告の効果は個人ごとに異なるが，その異質性は個人属性やブランドへの態度で予測可能．
2. 特にブランドへの態度の影響が大きく，ブランドに対して好意的であるほど広告効果は大きい傾向にあることがθkjから確認できる

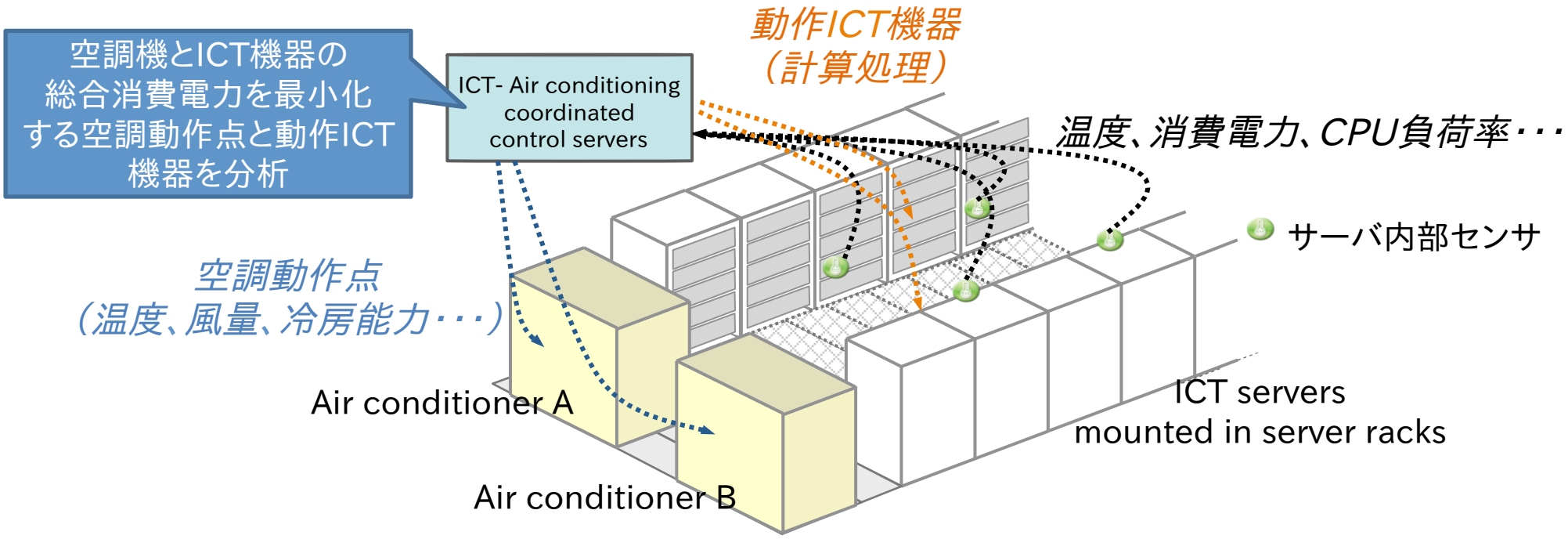
I-3

データセンタの省電力化に向けたICT機器吸込温度予測手法の検討

橋本英明，松尾啓吾 (日本電信電話株式会社)

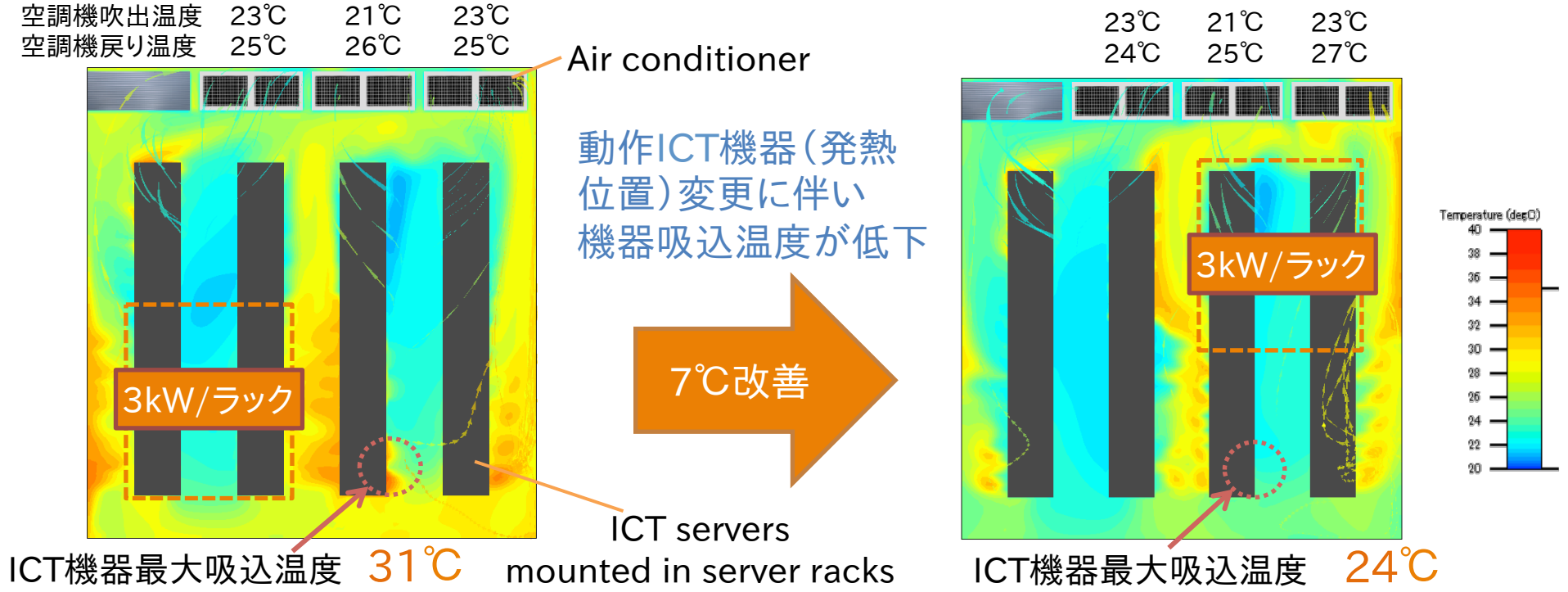
目的

- 空調機から冷却が容易なICT機器において，優先的に計算処理をすることで，空調機設定温度を緩和し省電力化を図る



橋本英明，松尾啓吾 (日本電信電話株式会社)

動作ICT機器の位置が冷却に与える効果



課題

- ICT機器の吸込温度や空調機の吹出温度等の運用情報に基づき，Dynamic PLSを用いてICT機器の吸込温度を精度良く予測する

I-4　動的残存効果モデルによる市場反応分析の高度化

井上 友彦（筑波大学大学院），佐藤 忠彦（筑波大学）

研究の目的

マーケティングにおいて，プロモーション活動量 (x_t) が売上 (y_t) に与える影響の経時的変化を構造化し測定する

構造上のポイント

- （プロモーションとは無関係の）ベースライン売上が存在する
- プロモーションは現時点だけでなく将来の売上にも影響する（効果の残存）
- ベースライン売上，プロモーション効果や残存率はすべて時間的に変化する
- プロモーション効果は正である

データ

- 製薬企業のディテール活動（人的販売）と製品売上の社内記録データ
- 週次，複数製品

I-4　動的残存効果モデルによる市場反応分析の高度化

井上 友彦（筑波大学大学院），佐藤 忠彦（筑波大学）

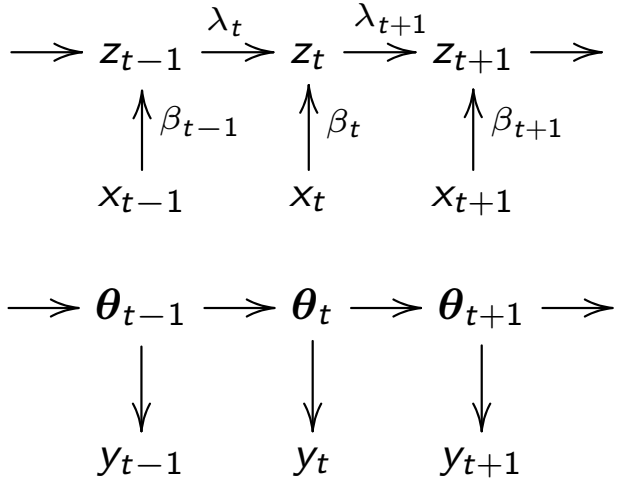
提案モデル（動的残存効果モデル）

$$y_t = \alpha_t + z_t + e_t, \quad e_t \sim \mathcal{N}(0, \sigma^2), \quad t = 1, \ldots, T$$

$$z_t = \beta_t x_t + \lambda_t z_{t-1}, \quad \beta_t > 0, \quad 0 < \lambda_t < 1$$

$$\boldsymbol{\theta}_t = (\alpha_t, \beta_t^*, \lambda_t^*)', \ \beta_t^* = \log(\beta_t), \ \lambda_t^* = \log(\tfrac{\lambda_t}{1-\lambda_t})$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + v_t, \quad v_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$$

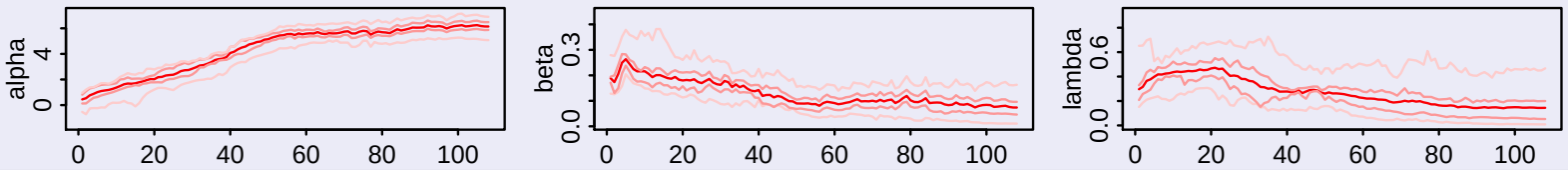


モデルの推定

粒子フィルタによる非線形・ガウス型状態空間モデルの推定

結果

プロモーション効果の経時的変化が妥当に推定され，製品毎の特徴も明らかに

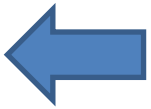


5 製品修理作業レポートと付随する数値データの関係性分析

山本忠,吉田稔,中川裕志(東京大学)渋谷久恵,前田俊二(日立製作所)

対象データ

no	document	cost index
1	set up new machine	6.123
2	transport new machine	6.112
3	remove cylinder no k03-38229	6.045
…	…	…
9674	replace oling of cooler	-2.165
9675	replace grease tank	-2.232



・メーカー製品の修理作業のレポート
  - 各文書は短い
  - 9675文書,3306語彙
・作業費用を示す
  コストインデックスが付随

分析

回帰分析：コストインデックスをより説明できるようなモデルを探す。
単語抽出：コストインデックスに大きな影響を与える単語を抽出する。

5 製品修理作業レポートと付随する数値データの関係性分析

山本忠,吉田稔,中川裕志(東京大学)渋谷久恵,前田俊二(日立製作所)

実行手法

	単語抽出	回帰
k-NN		○
Σy/f（各単語）	○	
LASSO	○	○
SVR （非線形）		○
SLDA	○	○
SVD＋回帰	○	○

・様々なモデルで回帰精度の比較や単語抽出をおこなった。

結果

1）回帰精度

手法	k-NN (k=5)	SVD+LM (dim:800)	LASSO (coef:665)	SVR	SLDA (topic:50)
MSE	0.7973	0.7972	0.7615	0.6496	1.0656

2）単語抽出

各モデルで単語の値段を算出する。（辞書作成）

I-7

購買履歴データを用いた消費者の選好構造の空間的表現手法の提案

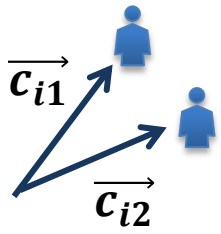
石田 実(アークエンジン)

提案手法

交互作用統計量は内積と解釈できる類似係数

購買の有無が2項分布に従うと仮定して、
消費者を空間的表現（理想ベクトルモデル）できる。

すなわち



消費者iのベクトル表現$\overrightarrow{c_i}$を下式とすると、

$$\overrightarrow{c_i} = \frac{1}{\sqrt{mp_i(1-p_i)}}\overrightarrow{m_i} - \sqrt{\frac{p_i}{m\,(1-p_i)}}\vec{e}$$

交互作用統計量　$s_{i1,i2} = (\overrightarrow{c_{i1}}, \overrightarrow{c_{i2}})$ 内積

ただし、消費者n人のm個の製品の購買履歴を表す行列を$\boldsymbol{M} = \{m_{i,j}\}, i = 1,\cdots,n; j = 1,\cdots,m$として、$p_i = \frac{\sum_k m_{i,k}}{m}$、$\overrightarrow{\boldsymbol{m_i}}$は$\boldsymbol{M}$の第$i$行、$\vec{e} = (1,\cdots,1)$

要検証

(1) この布置は購買の選好を表しているか？
(2) 新たな知見の発見ツールとして有用か？

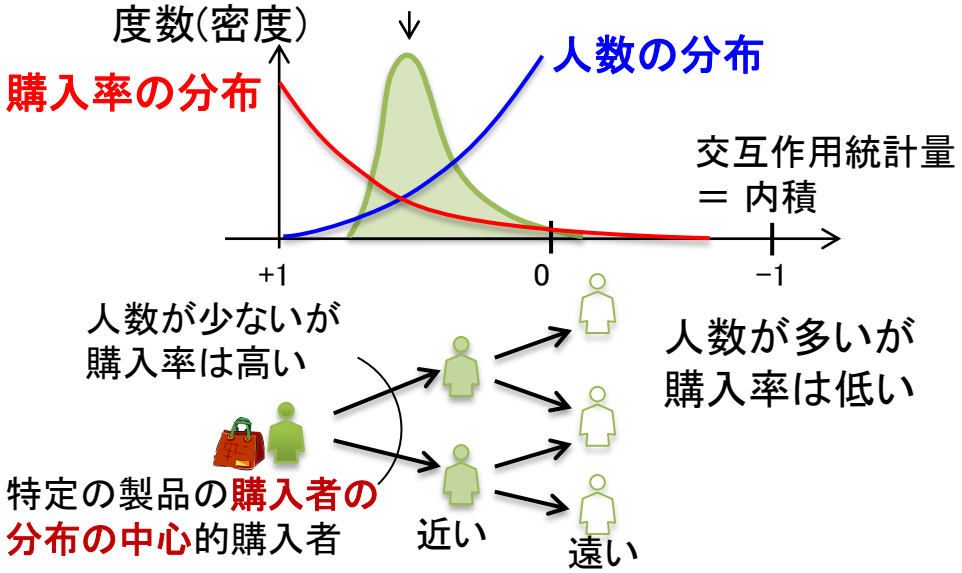
I-7

購買履歴データを用いた 消費者の選好構造の空間的表現手法の提案

石田 実(アークエンジン)

実証1 提案する布置は購買の選好構造を表しているか？ Yes

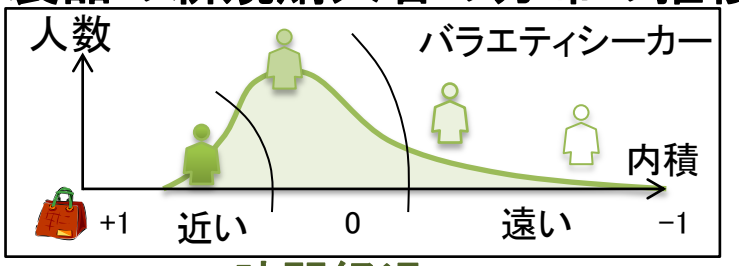
購入者数 = 人数の分布 × 購入率の分布



実証2 新たな知見の発見ツールとして有効か？ Yes

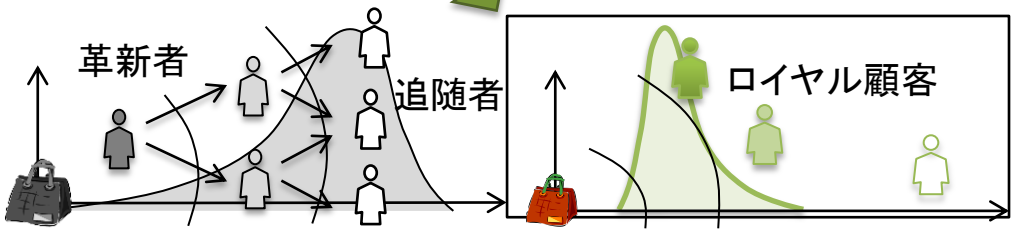
潜在顧客の分布の推移に関する知見を得た

新製品の新規購入者の分布の推移



時間経過

× 拡散(普及) ↘ ○収縮(同質化)



まとめ

提案手法は (1)線形表現なので、統計解析が容易。
(2)推奨システムや市場構造分析に有効。

Predicting Battery Life from Usage Trajectory Patterns

高橋俊博、井手剛
IBM東京基礎研究所

背景

- 電池は、使用するにつれ充電可能容量（容量維持率）が減っていく。
- 環境・使い方によって、劣化速度が異なる。
  - 劣化は、経時による劣化と、通電による劣化に分割できる。
  - 経時による劣化：下記パラメータによって、単位時間当たりの劣化速度が異なる。
    - 温度、SoC
  - 通電による劣化：下記パラメータによって、単位通電量当たりの劣化速度が異なる。
    - 温度、SoC、DoD

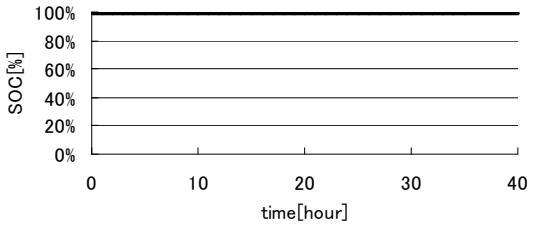
※
SoC（=State of Chageの略で、充電残量）
DoD（Depth of Dischageの略で、充放電時のSoCの幅）

課題

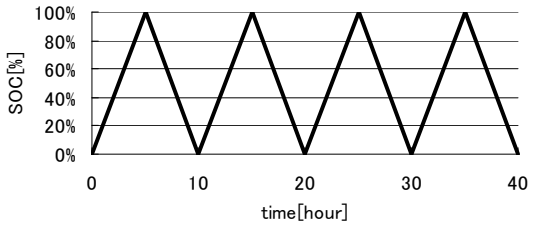
- 特徴量の設計をどうするか？
  - 支配的な劣化因子は定性的に分かっているが、パラメタライズの仕方は未だ議論されていない。
- 学習データが少ない
  - 予測モデル構築には、実際に劣化が進んだ多量のデータが必要。しかし、劣化試験はコストが高い。
  - 一方で、車の使い方は複雑（ブレーキ回生など）で、少ないデータを組み合わせて、これを当てる必要がある。
- 実車の実績データもモデル構築に用いたい

劣化試験パターン

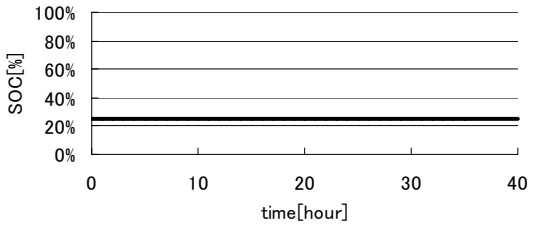
45℃、SoC100%、放置試験



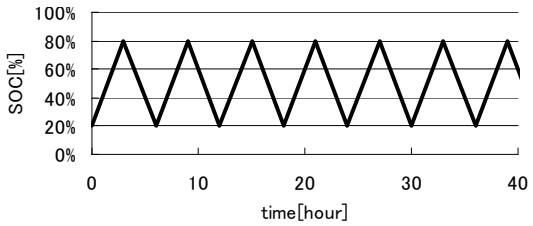
45℃、SoC0%-100%、DoD100%、サイクル試験



25℃、SoC20%、放置試験

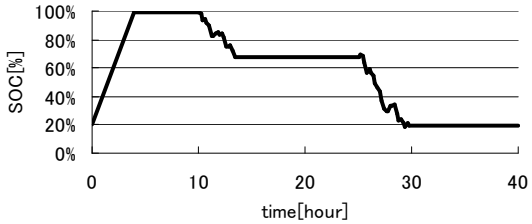


25℃、SoC20%-80%、DoD60%、サイクル試験



実走行パターン

実際の使われ方



限られた通電パターンの試験結果を組み合わせて、
複雑な通電パターンの電池劣化を予測したい。

Predicting Battery Life from Usage Trajectory Patterns

高橋俊博、井手剛
IBM東京基礎研究所

提案手法

- 「SoC × DoD × 温度」の3次元空間におけるトラジェクトリ回帰の問題として定式化

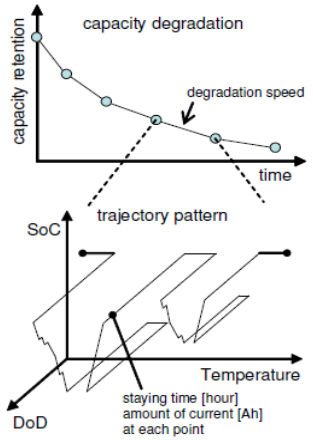
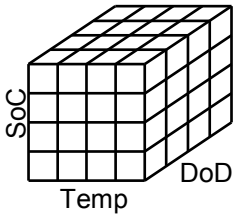


Figure 1. Outline of data set

離散化

「SoC × DoD × 温度」の3次元空間を離散化
各セルにおける滞在時間・通電量を積算する。
各セルに回帰係数を割り振る



劣化量を線形のトラジェクトリ回帰モデルで扱う

回帰係数　回帰係数

$$\phi(\boldsymbol{f},\boldsymbol{g}|\boldsymbol{\theta}) \equiv \sum_c \alpha_c f_c + \sum_c \beta_c g_c$$

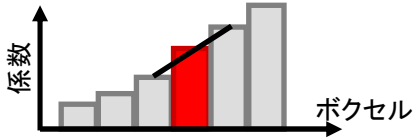
劣化量　セルcにおける滞在時間　セルcにおける通電量

残差項

$$L(\boldsymbol{\theta}) \equiv \sum_{n=1}^{N} \left(y^{(n)} - \phi(P^{(n)}|\boldsymbol{\theta}) \right)^2$$

- パラメータ数が多いので、自然な正規化を行う。

係数 α・β が滑らかに変化するように正則化
⇒隣接するボクセルの平均に近づける



正則化項

$$R(\alpha,\beta) \equiv \lambda_\alpha \sum_c \left(\alpha_c - \frac{1}{3} Q_{\mathrm{all}}(\alpha, c) \right)^2 + \lambda_\beta \sum_c \left(\beta_c - \frac{1}{3} Q_{\mathrm{all}}(\beta, c) \right)^2$$

目的関数 ＝ 残差項 ＋ 正則化項

$$\Psi(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$$

目的関数は二次関数なので容易に解ける

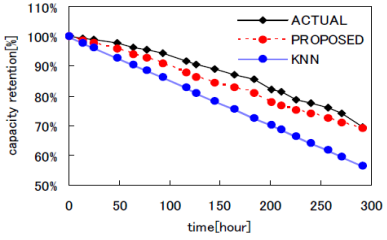
結果

- 実応用可能な電池劣化推定モデルを提案した。
- 実データを用いて、従来手法（kNNベース）と比較して、1.9倍～2.2倍の精度向上を達成した。

二乗誤差の比

	PROPOSED	KNN
SIMULATOR	1	2.20
ENDURANCE	1	1.89

予測の様子



I-9　医用画像におけるコンピュータ支援検出／診断のための機械学習：遠隔読影環境による多施設臨床使用下での識別器の更新

野村行弘、増谷佳孝、三木聡一郎、根本充貴、花岡昇平、吉川健啓、林直人、大友邦(東大病院)

CIRCUSシステム

- 病変自動検出をはじめとするコンピュータ支援検出／診断(CAD)ソフトウェアの研究開発／臨床応用促進を目的とした統合的な臨床情報処理基盤（2009.01より東大病院にて運用）

CIRCUS DB (DataBase)
- 機械学習のための疾患別病変データベースおよび登録システム（システム開発者向け）

CIRCUS CS (Clinical Server)
- Webインターフェイスによるマルチ CAD実行・評価サーバシステム
  ⇒ 評価（フィードバック）データはCADソフトウェアの性能評価・改善などに利用

※ CAD: computer assisted detection/diagnosis
CIRCUS: **C**linical **I**nfrastructure for **R**adiologic **C**omputation of **U**nited **S**olutions

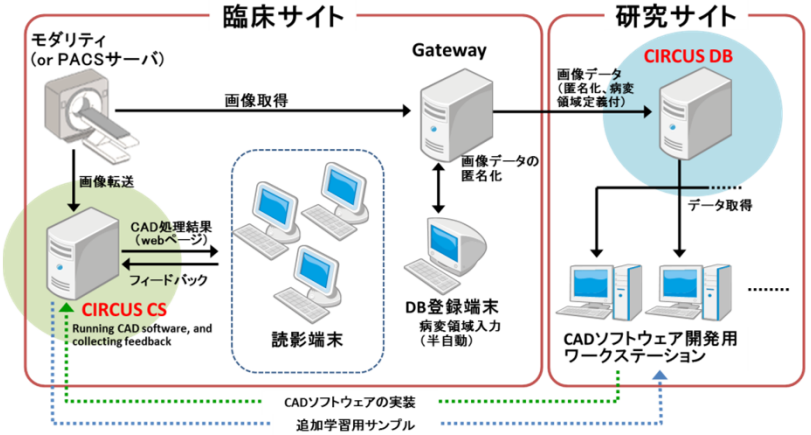


図1　CIRCUSシステムの構成

目 的

- 遠隔読影環境にてCIRCUS CSシステムを運用、多施設データに対するCAD実行およびフィードバックデータの収集
  - **CIRCUS+プロジェクト**（2011.09～）
  - 東大病院放射線科開発のCADソフトウェア（頭部MRA脳動脈瘤検出、および胸部CT肺結節検出）を使用
- 多施設データでの運用に伴う装置・撮像法の多様化による性能低下とフィードバックデータを用いた識別境界の更新による改善の定量化
  - 東大病院での学習結果をそのまま使用し、性能低下を確認。後にフィードバックデータを用いて再学習し、性能を再評価

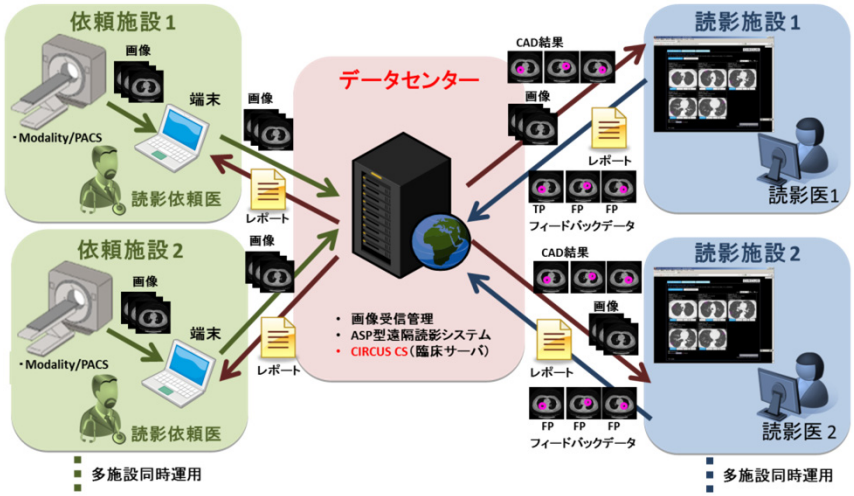


図2　遠隔読影環境の概要

方 法

- 東大病院データベースで学習したCADソフトウェアを多施設データに使用し、フィードバックデータを収集
- CADの更新は2種類のデータベース（東大病院・CIRCUS+）の情報を用いて偽陽性(FP)削減処理用識別器の再学習を行う

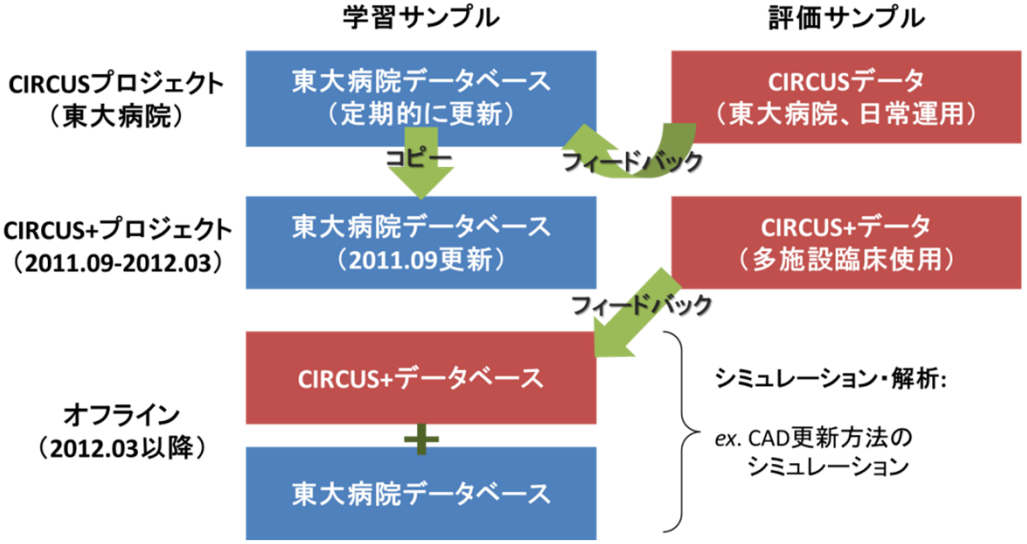


図3：CAD学習用データベースの更新

結 果

- 5施設による7ヶ月間の臨床使用で肺結節および脳動脈瘤検査の約6,000症例に対してCADの実行、およびフィードバックデータを収集
- 東大病院開発のCADソフトウェアを遠隔読影環境で使用した場合、性能低下を確認。フィードバックデータを用いた再学習により性能が改善（図4, 5）
  - 5 FPs/scanにおいて感度が7.4 %（肺結節検出）、8.1 %（脳動脈瘤検出）改善

まとめ

- 遠隔読影環境下にCIRCUS CSシステムを導入することにより、CADソフトウェアの多施設同時運用、およびフィードバックデータによる性能改善が可能

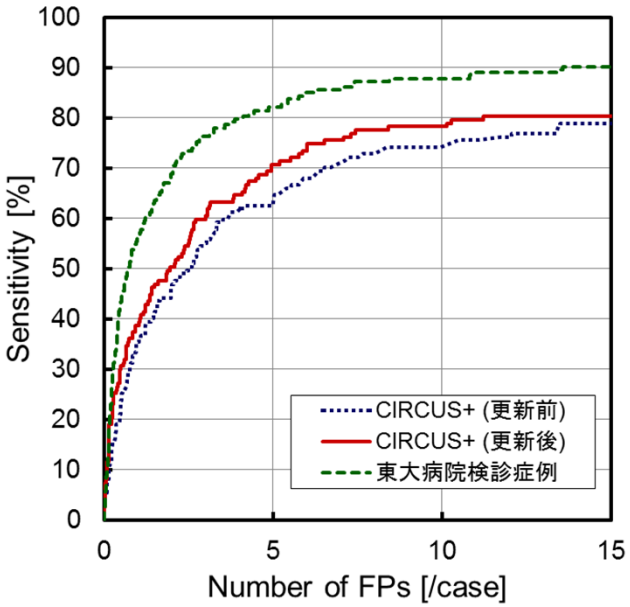


図4：性能評価結果（脳動脈瘤検出）

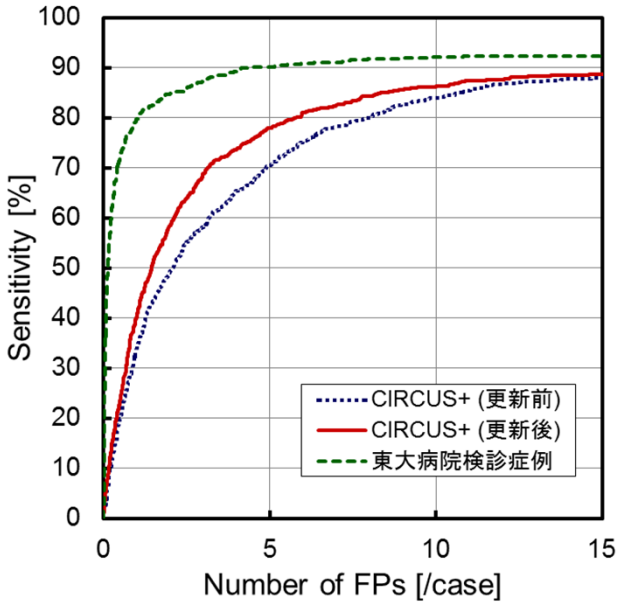
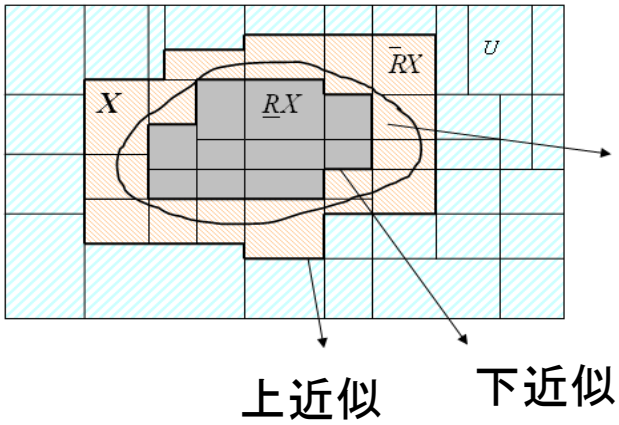


図5：性能評価結果（肺結節検出）

Rough setと部分空間法を組み合わせたリモートセンシング画像分類手法

ハスバガン, 山形与志樹 (独立行政法人国立環境研究所

ラフ集合（Rough Sets）



下近似 (Lower Approximation) と 上近似 (Upper Approximation)の計算

$$R_{\_}(X)=\{x\,|\,x\in U,[x]_E\subset X\}=\{x\,|\,x\in U,\forall y\in U[xEy\Rightarrow y\in X]\}$$

$$R^{-}(X)=\{x\,|\,x\in U,[x]_E\cap U\neq\varnothing\}=\{x\,|\,x\in U,\exists y\in U[xEy,y\in X]\}$$

訓練ピクセルxの下近似と上近似の計算

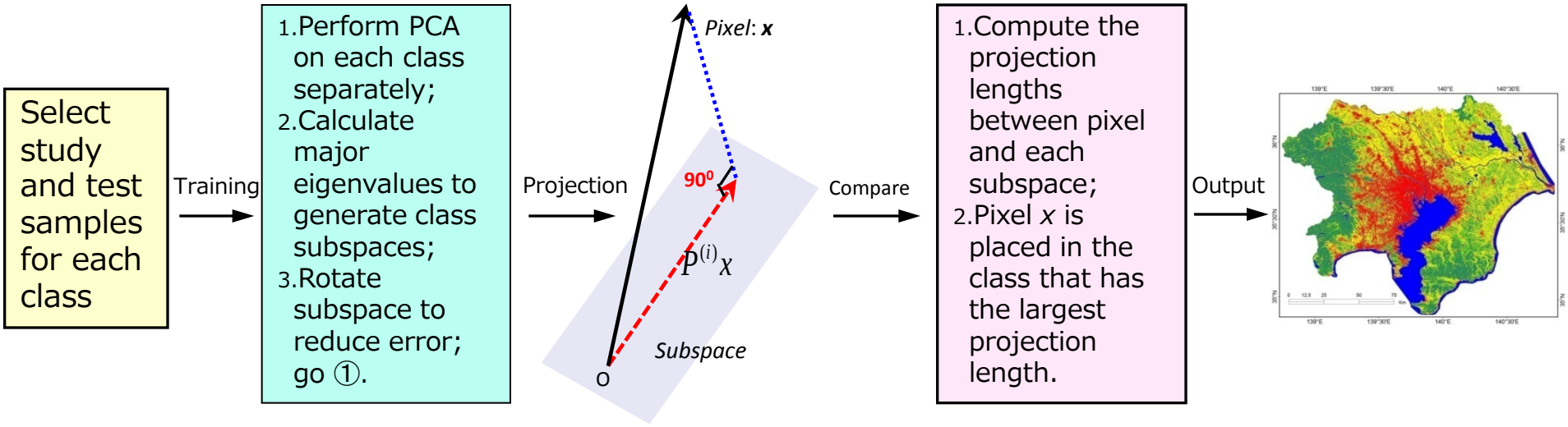
$$\underline{\tau_A}(x)\overset{\Delta}{=}\underline{\tau_A}(TS(x))=\{x\}\cup\bigcup_{y\in U,y\neq x}\{TS(y)\,|\,TS(y)\subseteq TS(x)\}$$

$$\overline{\tau_A}(x)\overset{\Delta}{=}\overline{\tau_A}(TS(x))=\{x\}\cup\bigcup_{y\in U,y\neq x}\{TS(y)\,|\,TS(y)\cap TS(x)\neq\varphi\}$$

訓練データを純化する

Step 1: 下近似で訓練データ xの各クラスでのメンバシップ値を計算し、メンバシップ値が高いクラスにxを配置する。曖昧の場合 Step 2;
Step 2: 境界集合でxの各クラスに所属する平均メンバシップ値を計算し、値が高い方に配る。まだ曖昧の場合， xを訓練データ集合から削除する;
Step 3: 配置された全ての訓練データを精査し、元のクラスと違うクラスに配置された訓練データを削除する。

純化したデータを使った部分空間法



I-11　創薬を支援するデータ駆動型化合物設計

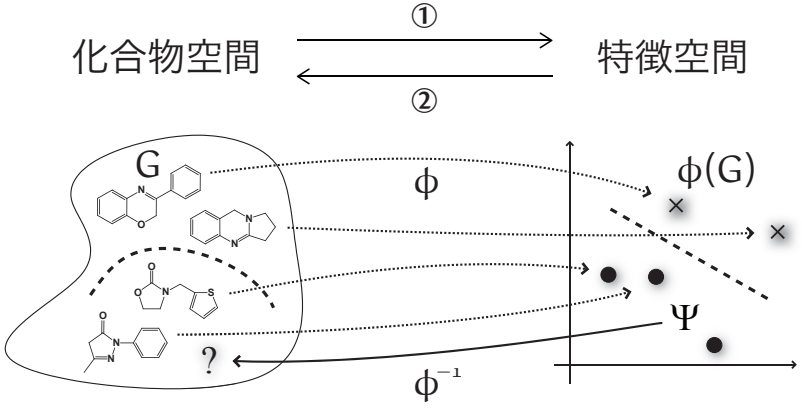
山下博史　総合研究大学院大学
吉田亮　統計数理研究所
伊庭幸人　統計数理研究所
樋口知之　統計数理研究所

□ 問題：化学構造の非凸・非線形・組合せ・多目的・最適化

創薬：薬に必要な機能を複数併せ持つ化合物を広大な化合物空間から実験を通して試行錯誤的に探索するプロセス

□ 目的：データから目的の機能を持つ化学構造を推定する
(グラフpre-image探索手法の開発)

提案手法



① 化学構造の機能予測に適した特徴空間の構成

✏化学構造用カーネルの設計

② 特徴空間で定義した目標分布からの化学構造サンプリング

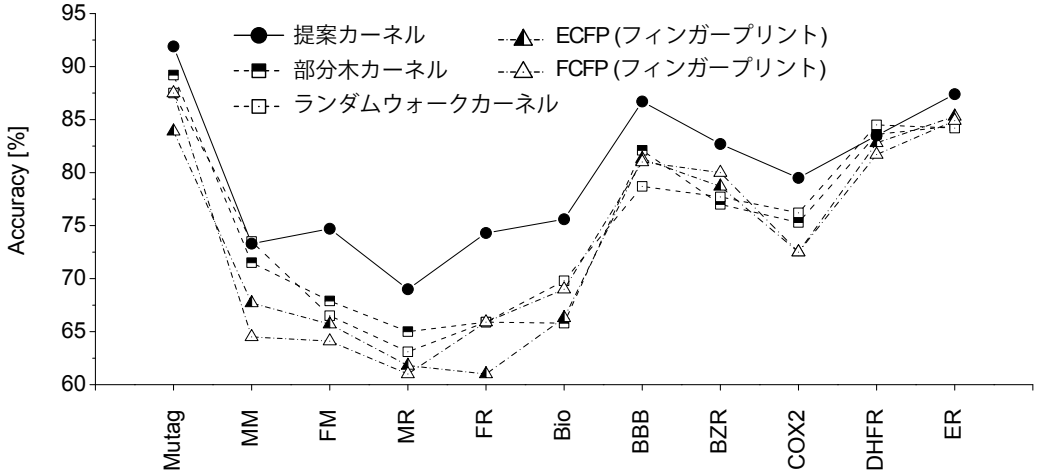
✏MCMCにおける提案分布の設計

化学構造の分類実験

提案カーネルの性能評価(10-fold cvにより予測精度を計測, 学習器にはSVMを使用)

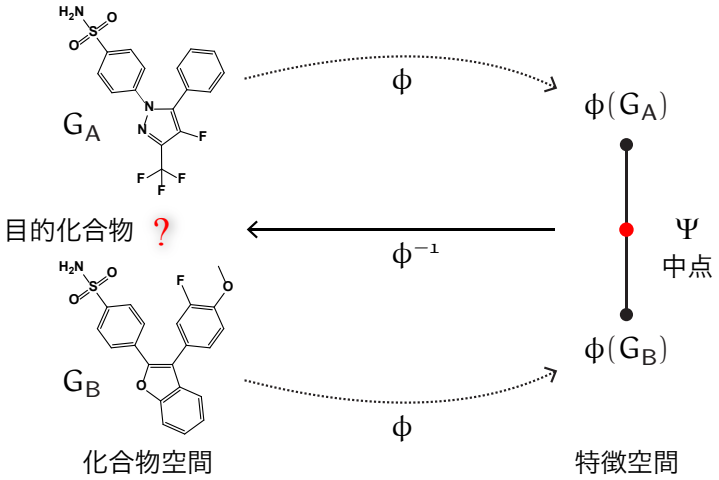
データセット

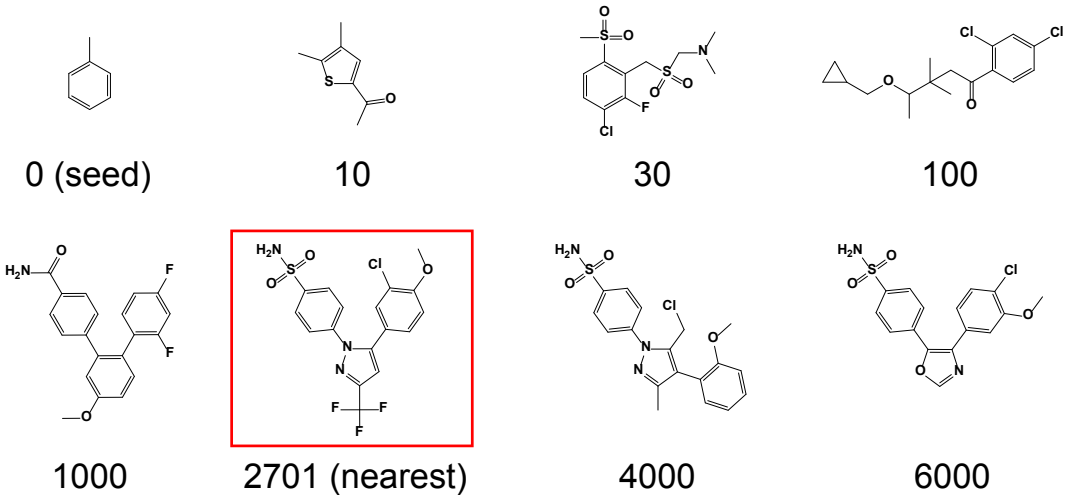
変異原性	Mutag
発ガン性	MM, FM, MR, FR
生物学的利用能	Bio
BBB透過性	BBB
タンパク質結合能	BZR, COX2, DHFR, ER



グラフpre-image探索実験

特徴空間の点Ψに埋め込まれる化学構造をMCMCを用いてサンプリング





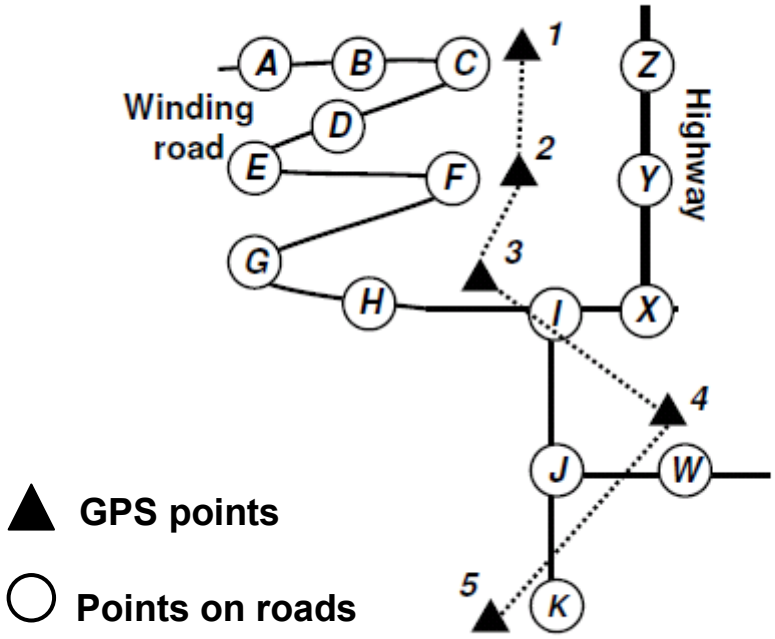
MCMCシミュレーションにおける化学構造の推移

I-12 An Online Map Matching based on Hidden Markov Model

Rudy Raymond, Sei Kato, Tetsuro Morimura (日本IBM), Masato Hattori(青山学院大学)

Goals of online map matching:

Given a sequence of GPS points and a map of road network, *find the sequence of roads* that most likely produce the points from the map *online*



Possible road sequences from (1,2,3,4):
A-B-C-D-E-F-G-H-I-J-K
Z-Y-X-I-J-K
…

Why important?

- Prerequisite to finding patterns in mobility with efficient resources
- The base for many algorithms, e.g., in a traffic simulation that extracts parameters from probe car data

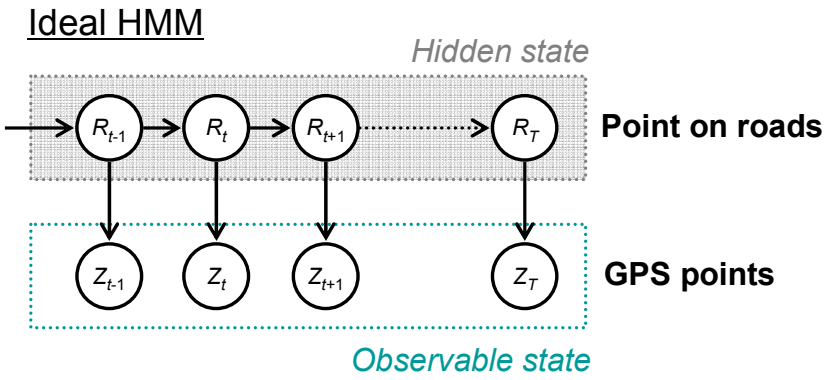
I-12 An Online Map Matching based on Hidden Markov Model

Rudy Raymond, Sei Kato, Tetsuro Morimura (日本IBM), Masato Hattori(青山学院大学)

Our approach:

HMM-based online map matching

- GPS are observed states and roads are hidden states
- Memorize the current state only
- Moderate computational cost
  - -- Simple emission and transition probabilities
  - -- Heuristic for storing potential matching candidates



Method	Frechet Dist.	Avg. Frechet Dist.
Simplified HMM	1.428	0.255
Online HMM	1.554	0.269

Table 1: Comparison on TT dataset

Results:

- Online map-matching without time-sliding windows (delay)
- Comparable accuracies with the offline version
  - -- Real-world sparse and noisy datasets

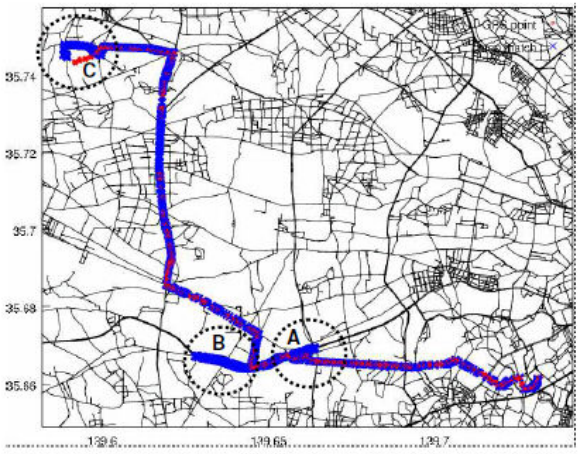
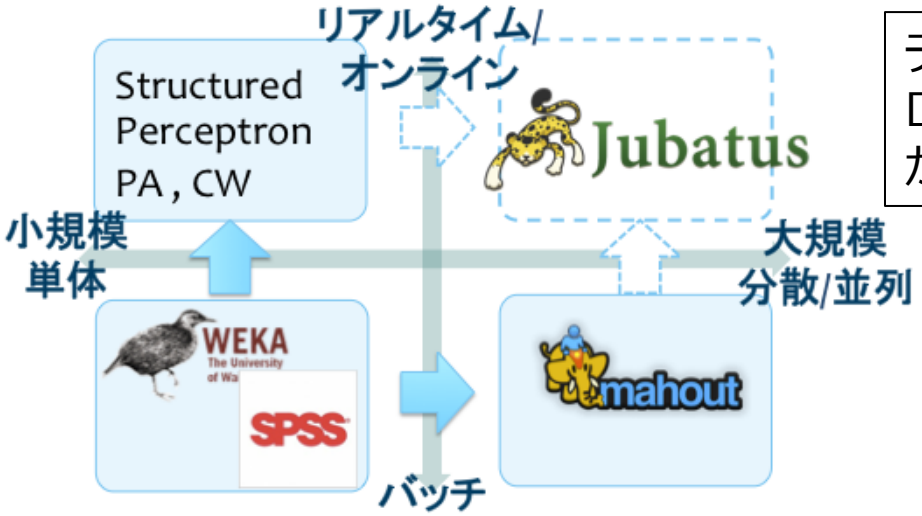


Figure 5: A map matching sample on TT

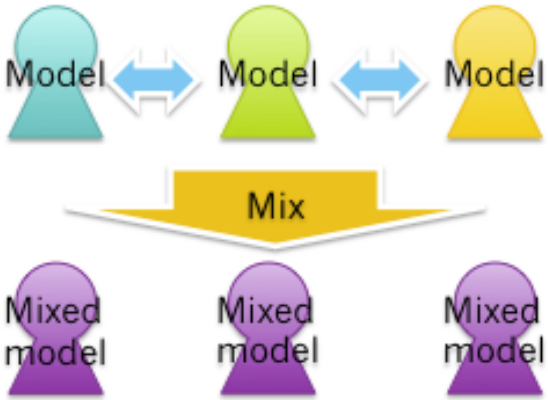
Jubatus:
大規模データ解析向け分散オンライン機械学習

大野健太　海野裕也　岡野原大輔　比戸将平
株式会社 Preferred Infrastructure



データ増加に対抗する機械学習のアプローチは分散化、オンライン化があるが、Jubatusは2つの両立を特徴とする。

分散オンライン機械学習では頻繁な更新と同期を同時に達成しなければならない



JubatusはUpdate-Mix-Analyzeという仕組みにより学習モデルを緩やかに共有する事で分散オンライン機械学習を実現している。

I-14

Modeling Patent Quality:
A System for Large-scale Patentability Analysis using Text Analytics

鈴木祥子（日本アイ･ビー･エム株式会社）

提案手法

特許の質を評価する指標を導入するため、教師あり学習を行い各特許明細書のスコアを計算
- -明細書から各種特徴量を抽出し、数値化
- -教師データとして、審査請求された特許が成立するかどうかというラベルを利用

特徴量

- •明細書の各種統計量
  - - タイトルの文字数、請求項の数、etc
- •テキスト解析を利用した特徴量
  - - 形態素解析･係り受け解析を利用した構文複雑性の導入
  - - TF-IDF
  - - 単語年齢の導入

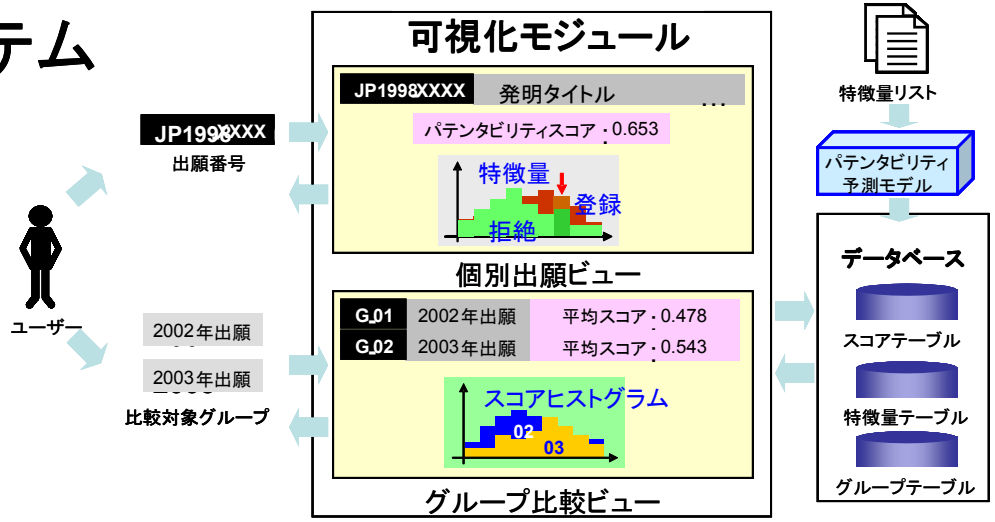
学習モデル

ロジスティック回帰モデル
- - L2正則化
- - 明細書の出願年についてマルチタスク学習を適用

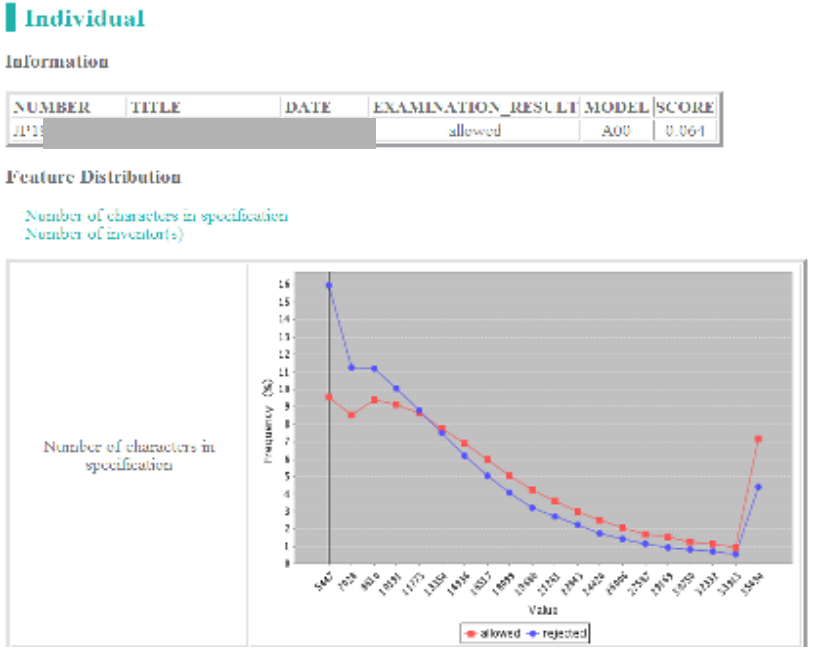
I-14 Modeling Patent Quality: A System for Large-scale Patentability Analysis using Text Analytics

鈴木祥子（日本アイ･ビー･エム株式会社）

システム



システムアーキテクチャー



可視化モジュール出力例

予測精度

IPC=H01l : AUC=0.62
IPC=G06f : AUC=0.69

まとめ

- 特許の質を客観的な指標で評価するシステムを500万件超の明細書で実現
- 予測精度はまだ十分でないが今後の工夫で精度向上の可能性が高い

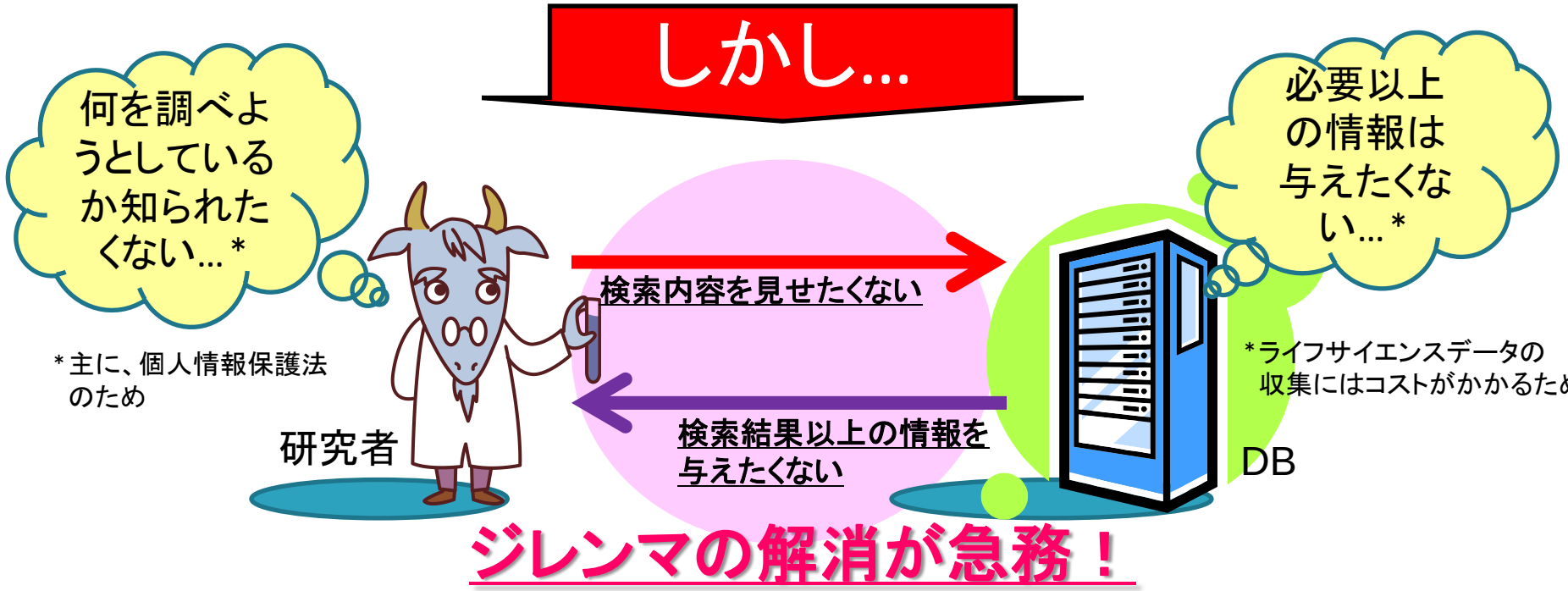
I-15

化合物データベースの秘匿検索技術の開発

清水佳奈[1], 荒井ひろみ[3], 縫田光司[2], 浜田道昭[5], 津田宏治[1], 広川貴次[1],
花岡 悟一郎[2], 佐久間淳[4], 浅井 潔[5]
1．産総研・CBRC, 2．産総研・RISEC, 3．理研, 4．筑波大, 5．東大

背景

- 近年におけるライフサイエンスデータの目覚ましい充実化
  - ゲノムデータ，タンパク質相互作用データ，糖タンパク質データ，…
  - これらのデータベースを全面的に活用することにより，医学・生物学の著しい発展が期待される．



本研究で取り組んだ課題

- 創薬研究者が有料化合物DBの購入を検討する際には，興味のある化合物と類似する化合物がＤＢ中にいくつ含まれているかを知ることが重要.
- ユーザー側とサーバー側が互いに情報を秘匿したまま，ユーザー側のみが，クエリと類似する化合物の個数を知る技術を開発した.

提案手法

$$S_{Tanimoto} = \frac{|p \cap q|}{|p \cup q|} \geq \frac{\theta_n}{\theta_d}$$

$$\leftrightarrow (\theta_d + \theta_n)|p \cap q| + \theta_n(-|p|) + \theta_n(-|q|) \geq 0$$

- 上式の左辺をTとする．化合物 p, q の類似度が閾値以上の場合，T は非負となる.
- **加法準同型暗号**を用いてT を計算し，ユーザー側のみが結果を得る.
- DB中の化合物全てに関して同様の計算を行えば，ユーザー側のみが類似化合物の個数を得る事ができる.

実験結果とまとめ

- 化合物の秘匿検索技術を開発した.
  - 従来技術（ＭＰＣ）と比較して，計算コスト，メモリ使用量が大幅に低下.
  - 必要とされる通信は1往復.
- ライフサイエンス分野の様々なデータへの応用を検討している.