

IBMにおける機械学習研究と応用

IBM東京基礎研究所 数理科学担当

井手 剛





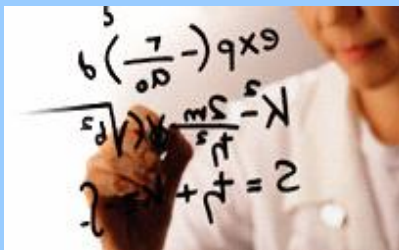
Contents

- IBM Researchの概要
- IBM Research数理科学部門の紹介
- IBM東京基礎研究所・数理科学部門の紹介
- プロジェクト事例: トラジェクトリ回帰とその応用
- まとめ



IBM研究部門のストラテジー・エリア

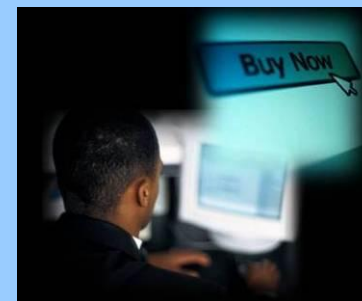
数理科学



インダストリー・ソリューション



サービス



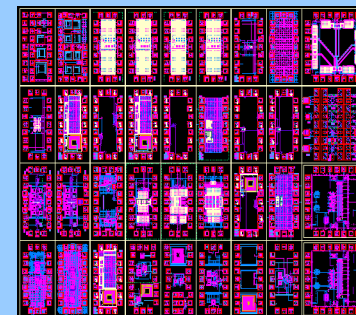
ソフトウェア



システム



テクノロジー



基礎科学研究



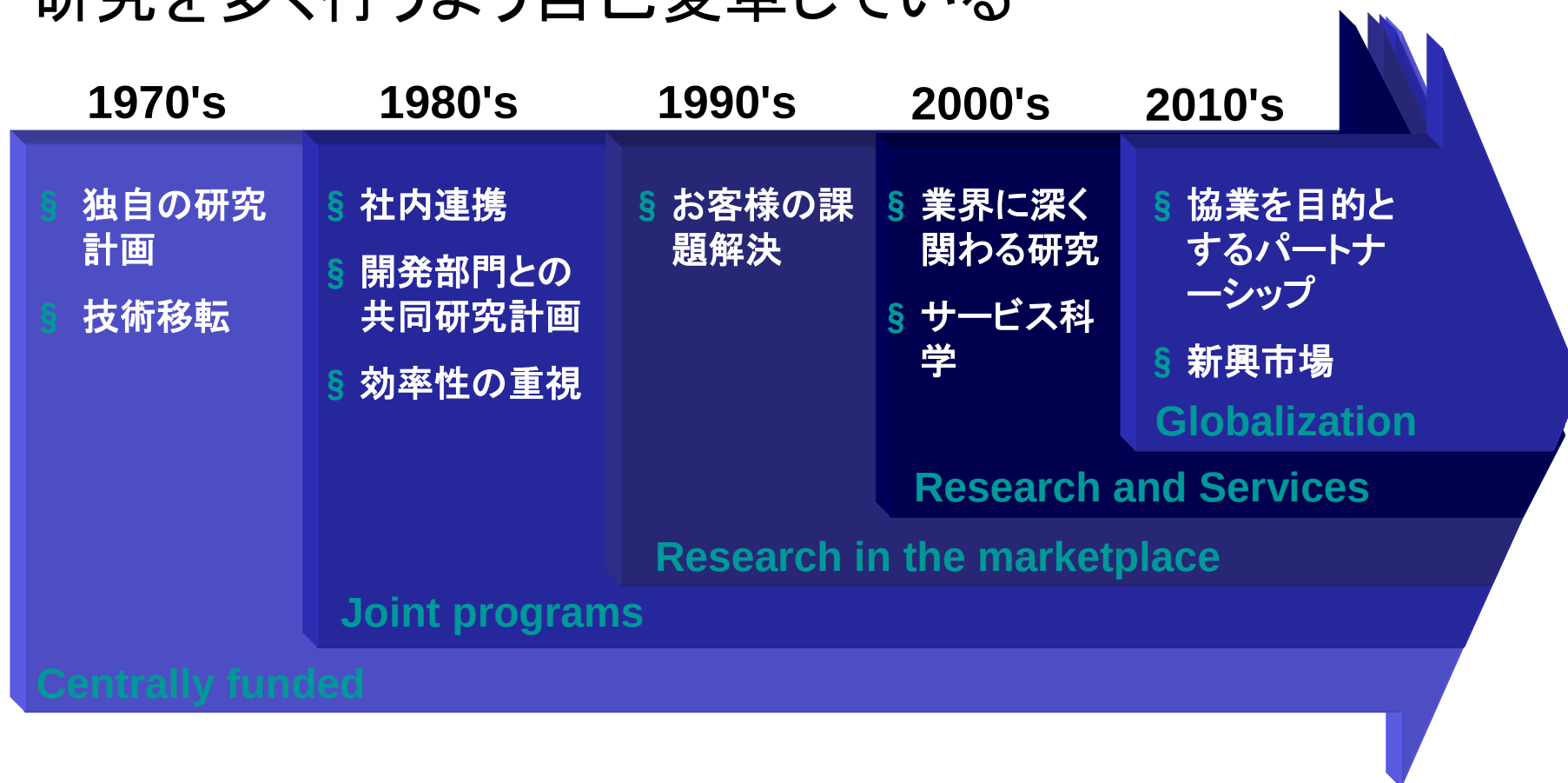


ここ10年でIBMの研究所の数は、新興国を中心に倍増している。今年はアフリカにも開設された





IBM研究部門は、外部のパートナーとの開かれた基礎研究を多く行うよう自己変革している



Inter-disciplinary collaboration in the market and across the globe

お客様プロジェクトへの参画は、数理解析技術をベースにしたサービスを行う専門のコンサルタント部隊と協業するのが通例

- ルーティン業務の自動化はすでにコモディティ化している
- 高度な解析技術に基づいて、ビジネスに有用な知見を与えることが求められている
- IBMでは、SPSS、CPLEX等のツールを元に解析技術のトレーニングを受けたコンサルタント部隊を育成している
- 典型的な数理解析の問題はBAO (business analytics & optimization) と呼ばれるコンサルタント部隊がで処理する
- 定式化されていない問題、困難な問題は研究部門と共同で解決に当たる



Japan [変更]

検索

ホーム ソリューション サービス 製品 サポート & ダウンロード My IBM

ようこそ [ログイン] [登録]

「BAO」が導く新たな知見。 膨大な情報を経営に生かす秘訣とは？

散在する多様な情報からビジネス・チャンスを発掘。
予測的分析に基づく業務最適化で、さらなる成長を牽引します。

はじめに BAOでビジネス・チャンスを ソリューション お客様導入事例

BAOソリューション体系
ビジネスの課題にきめ細かく対応させ、BAOサービスを利用しやすくしました。
ソリューション体系図を表示する

情報活用診断
膨大なデータをもっと有効活用したいが、どこから手をついたらいいのかというお客様に最適です。情報活用に関するお客様の現状を調査・診断して、あるべき姿と現実のギャップを明らかにし、優先着手領域を特定します。
・情報活用の成熟度診断
・BIアプリケーション診断

クイック・アセスメント・サービス
「ソリューションの本格導入の前に効果を検証したい」というお客様のご要望に対して、短期間に効果を検証します。
・KPI定義による見える化構想支援
・テキストマイニングを活用したお客様の声活用
・異常値の効率的な検知
・統計アプローチに基づく在庫最適化
・価格最適化

BAOソリューション

- スマートな経営の可視化**
 - ・ビジネス・パフォーマンスを高める経営の可視化
 - ・企業の活動評価指標の最適化
- スマートな顧客分析**
 - ・お客様の声活用
 - ・インターネット上の顧客行動の分析
 - ・販促費用の最適化
 - ・販売価格の最適化
- スマートな生産と物流**
 - ・配送経路の最適化
 - ・生産計画と製造ラインのスケジューリング
 - ・動的在庫の最適化
 - ・製造現場と経営をつなぐマネージメント・コックピット
- スマートな品質管理・異常検知**
 - ・異常値の検知
 - ・映像・画像の自動分類と検索
- スマートなコンテンツ管理**
 - ・ワンソース・マルチユースを実現するドキュメント作成

フォームお問い合わせ
まずはお気軽にご相談ください。
入力フォーム

IBM Smart Analytics System
スマートなビジネスを加速する分析アプライアンス
詳細はこちら

最先端のビジネス・アナリティクス



Global Technology Outlook (GTO): 研究部門は、会社全体のビジネスの方向性を毎年答申する

□GTO 2012のテーマにはAnalyticsが関係するものが多い

–おそらく今はAnalyticsの黄金時代

Managing Uncertain
Data at Scale

Systems of People

Outcome Based
Business

Resilient Business
and Services

Future of Analytics

The Future Watson



Smarter Planet: 次世代社会基盤の構築をIT主導で進めるためのIBMの戦略的ビジョン。Analyticsはそこで主要な役割を果たす



□ ITインフラの成熟は客観的に明らか

□ 高付加価値のビジネスは、従来のIT基盤の上で行われる何らかのインテリジェントな処理

□ IT基盤と伝統的な社会基盤を、IT主導で融合することで巨大なビジネスインパクトを目指す

□ IT主導 = Analytics主導



Contents

□IBM Researchの概要

□IBM Research数理科学部門の紹介

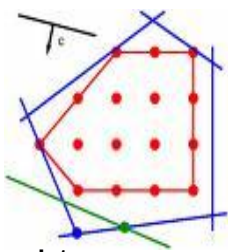
□IBM東京基礎研究所・数理科学部門の紹介

□プロジェクト事例: トラジェクトリ回帰とその応用

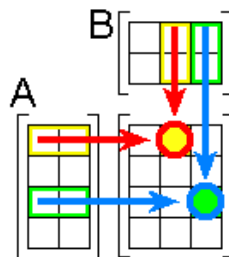
□まとめ



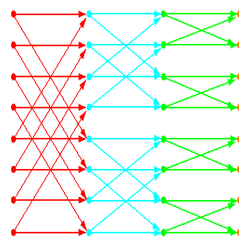
IBM Researchの数理科学部門は、過去40年以上にわたり重要な研究成果を生み出してきた



Integer Programming

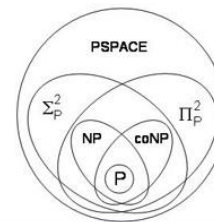


Fast Matrix Multiplication

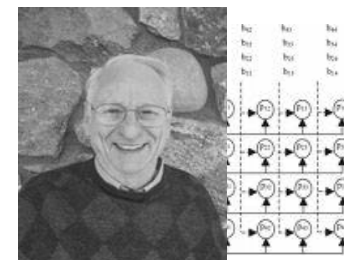


Fast Fourier Transform

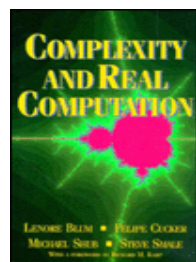
Parallel computing



Foundations of Complexity



Combinatorial Matrix Theory



Complexity of reals



Lattice based Cryptography



Adversarial Queueing

Data Encryption Standard



Quantum FFT



Trackpoint



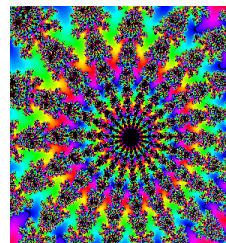
Algebraic Complexity



Priority Encoding



Topological Entropy



Fractals



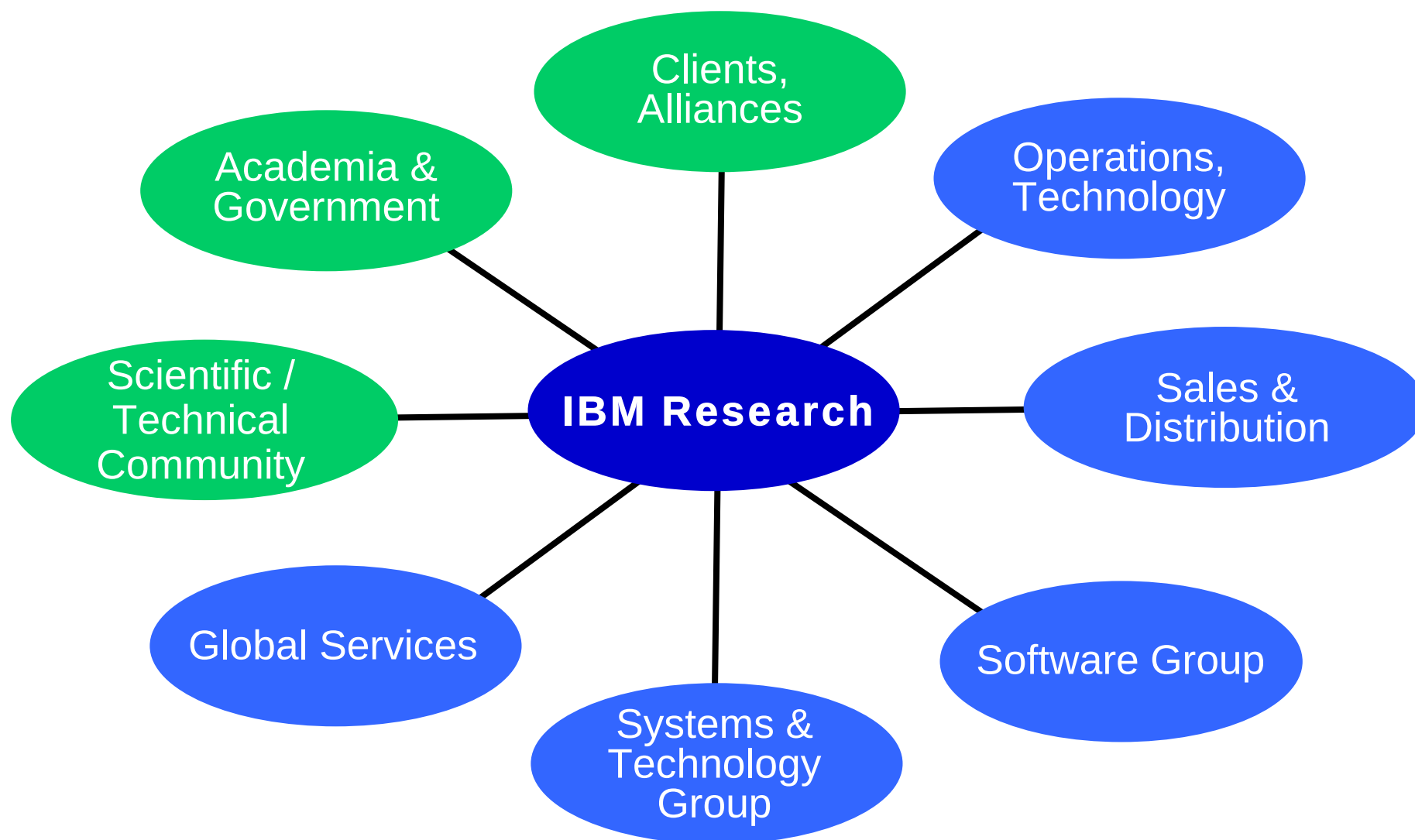
Descriptive Complexity



Volume Algorithm



製品およびサービスの差別化のため、数理科学部門に寄せられる期待は非常に大きい





IBMは、Analyticsのビジネス応用に関するほぼすべての分野をカバーしている



出典： Competing on Analytics, Davenport and Harris, 2007

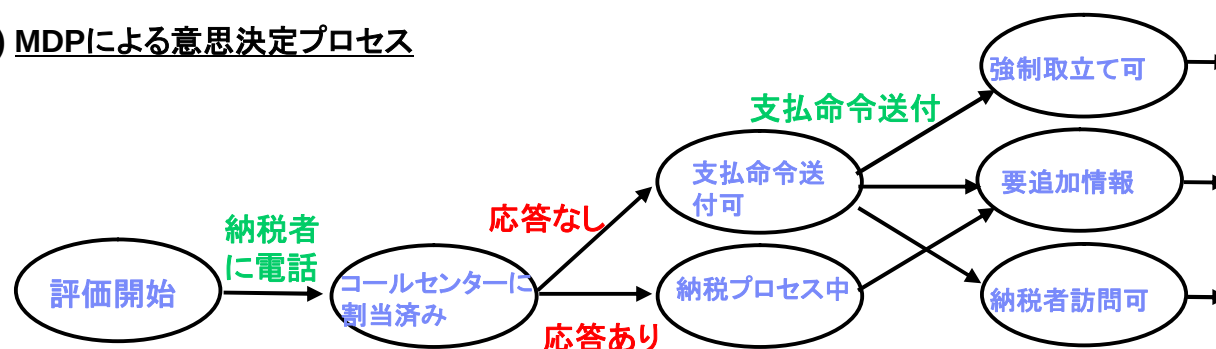
数理科学部門が関係する最近の主要プロジェクト

TACOS – IBMの徴税支援システム

KDD 2010 Best Paper

マルコフ決定過程(MDP)を用いて、納税申告書の異常を判定し、疑わしい納税者へのアクション(電話/手紙/訪問)を推薦する。米国のNY州政府はこのシステムにより、3年間で1億ドルもの巨額の追加税収を得る見込み

例) MDPによる意思決定プロセス



Deep QA – クイズ番組でチャンピオンを目指す

Open-domainの質問応答システム
DeepQAが、「ジョパディー」というアメリカの
クイズ番組で最高のチャンピオンに勝利。

最先端の自然言語処理技術と高度な数理
解析技術がカギ。

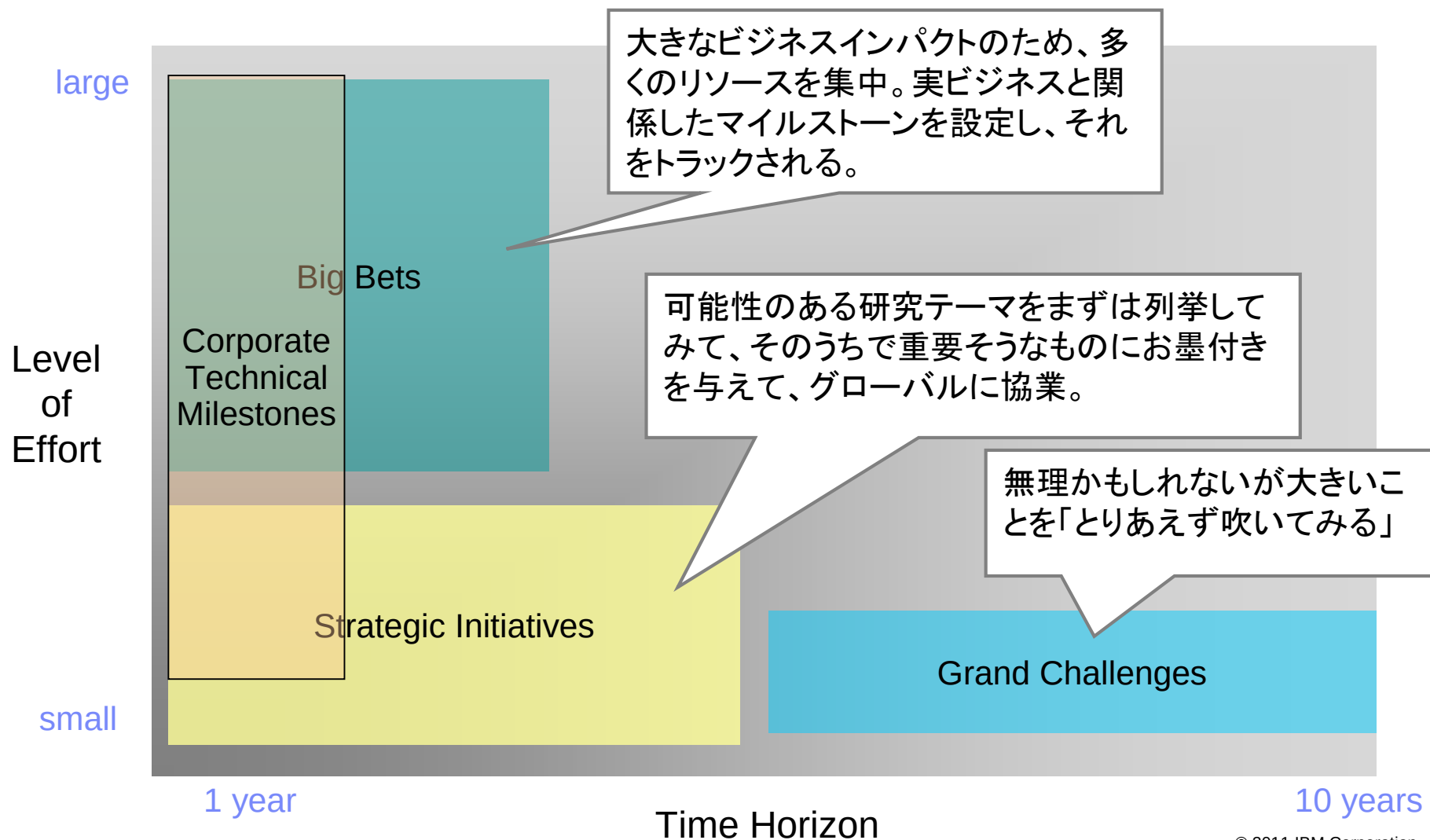
Watson takes on Jeopardy!

Advanced computing system
has potential to take business
intelligence to a new level





IBM研究部門における重要研究テーマのカテゴリ





Contents

- IBM Researchの概要

- IBM Research数理科学部門の紹介

- IBM東京基礎研究所・数理科学部門の紹介

- プロジェクト事例: トラジェクトリ回帰とその応用

- まとめ



IBM東京基礎研究所の数理科学部門の専門分野

機械学習

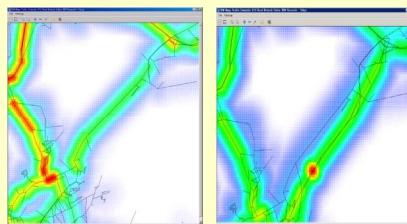
最適化とアル
ゴリズム

シミュレー
ション

最近の代表的なプロジェクト

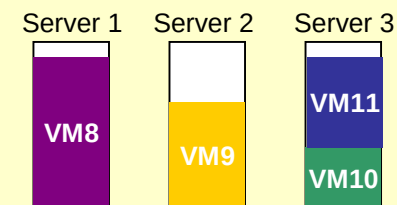
Megaaffic (IBM Mega Traffic Simulator)

- 超大規模のマルチエージェント交通シミュレーター
- 運転手の行動パラメータの自動学習機能を世界で初めて搭載



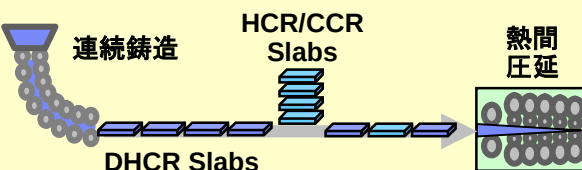
DVA (IBM Dependable VM Allocator)

- バーチャルマシン (VM) を複数のサーバーに割り当てる仕方を動的に最適化
- 手作業に比べて30%以上のエネルギー効率を達成



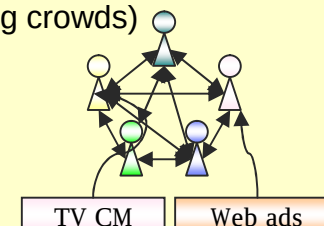
PDOS: 最適鉄鋼生産スケジューリング

- 世界最新鋭の製鉄所で圧倒的シェアを誇る生産最適化ソリューション



SimPHOMICS (IBM Simulation and prediction for what-if marketing with interacting crowds)

- エージェント・シミュレーションを用いて、口コミ等の伝播過程を計算
- いつどういう広告を打つべきかのWhat-if解析を行う

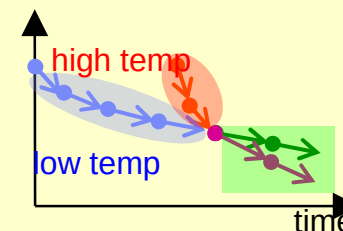


Anaconda (IBM Anomaly Analyzer for Correlational Data)

- 変数の依存関係を解析し異常の兆候を察知する異常解析エンジン
- 鉄道、自動車、船舶等での多数の事例

QUINTET (IBM Quick analysis toolkit for energy saving technologies)

- 電池を用いた省エネルギー技術のための解析スイート
- 電気自動車の電池寿命の予測誤差を数分の1に改善





最近の重点領域

複雑に相互作用するシステム
の最適化

意思決定の
科学

センサーデータ
解析

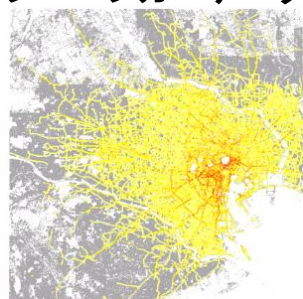
複雑に相互作用するシステムの最適化

- エージェント・シミュレーションは、人や物が複雑に相互作用するシステムを多様なシナリオにおいて評価するための有力な手法
- シミュレーション結果はモデルのミクロなパラメータの値に強く依存しますが、多くの場合、ミクロなパラメータは直接観測不可能であることが多く、マクロな観測量からミクロなパラメータを推定する技術が必要
- たとえば、交通シミュレーションの文脈で、実用的なモデル推定技術を開発している

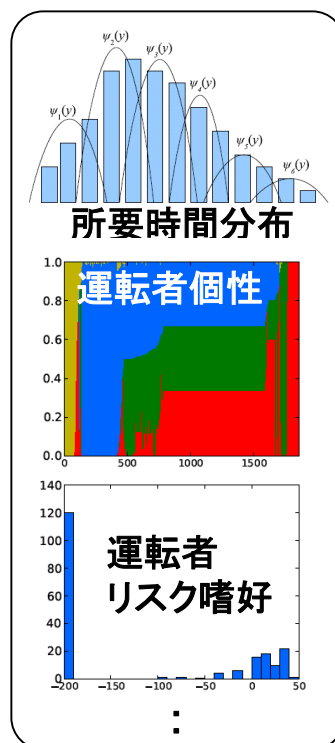
正確なシミュレーション・モデル構築

複雑に相互作用するシステムの最適化

プローブカーデータ

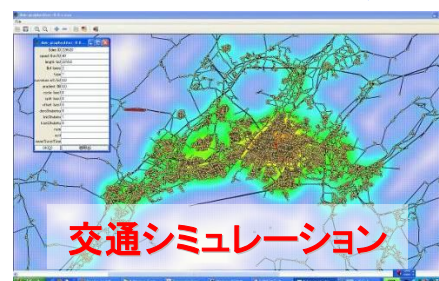


マクロ観測量から
ミクロパラメータ
推定



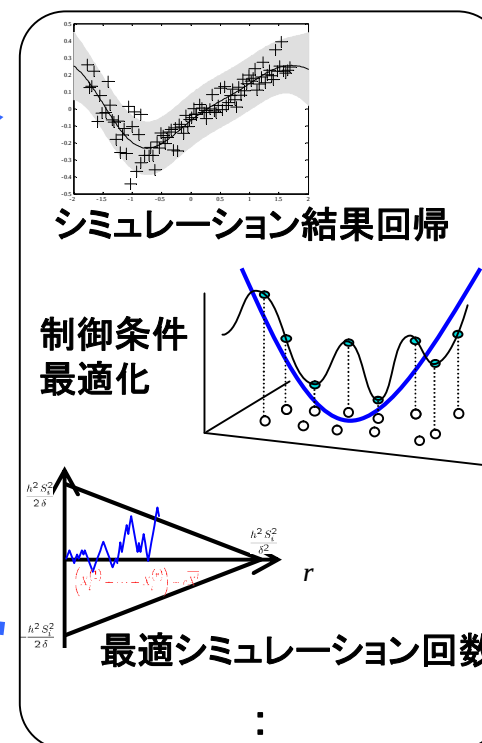
パラメータ

結果



パラメータ
修正

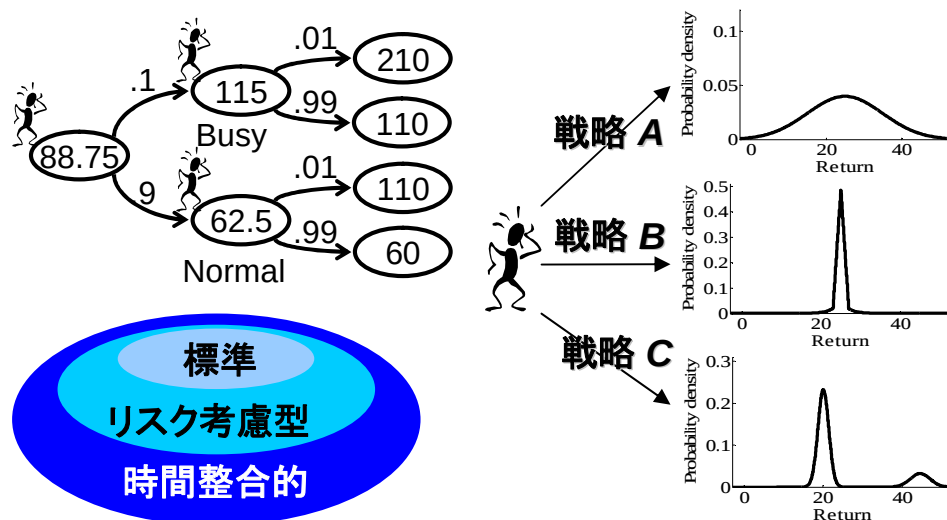
制御条件



意思決定の科学

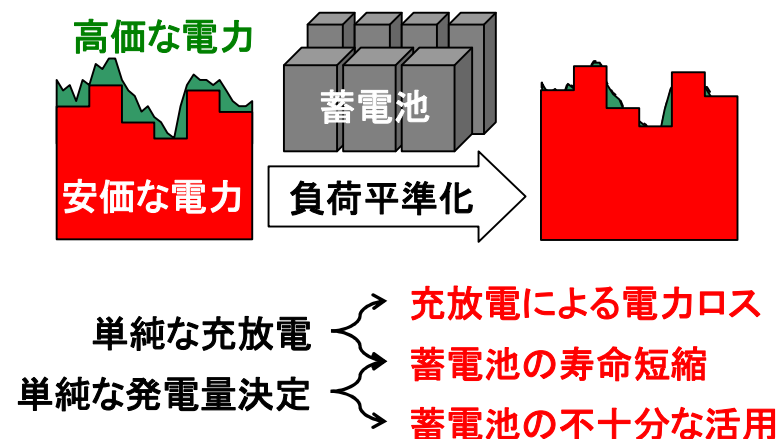
- ビジネスにおける意思決定を合理的かつ最適なものにするために、数理科学に基づく最適化技術を開発している。
- 特に、(1)逐次意思決定、(2)リスク考慮型意思決定、の理論面、応用面の研究を進めている

多様な合理的意思決定を支える マルコフ決定過程・強化学習

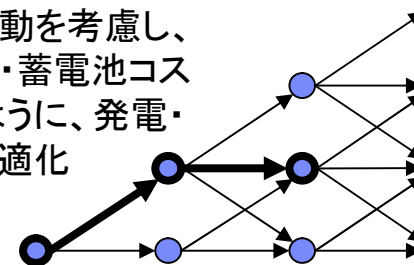


不確実な情報に基づく意思決定を最適化するマルコフ決定過程・強化学習技術において、考慮できる合理的なリスク嗜好が限られているという課題を解決

応用例: 発電・充放電計画の最適化



不確実な需要変動を考慮し、中長期的な電力・蓄電池コストを最小にするように、発電・充放電計画を最適化





センサーデータ解析

- 製造業などで取得されるセンサーデータは、強いノイズ、複雑な相関構造、など解析困難な特徴を持っている。最新の機械学習の技術により、これらの困難を突破するための研究を行っている

設計パラメーター最適化

- 過去のシミュレーションの結果を元に、任意の設計パラメーターにおける性能を予測
- 設計開発の工数短縮への有用性を確認

自動車組立工の作業ミス監視

- 作業員の身体につけた地磁気センサーを元に、動作異常を識別
- 実データを使った検証で良好な精度を確認

鉄道車両状態監視

- 車軸の温度をモニタリングし、異常を早期検出
- 従来手法より数倍高い検出能力を実証

自動車生産ラインの品質監視

- 溶接時に出る閃光のスペクトルを分析し、良・不良を瞬時に判別
- ほぼ100%の判別精度を達成

船舶の状態監視

- 海象の影響により船舶主機の状態監視は困難
- 最新の異常検知技術で専門家の診断を上回る性能を実証

発電設備状態監視

- 中規模の発電設備に設置された100以上のセンサーを監視
- 人の目で判別が困難な異常の早期検知に成功



Contents

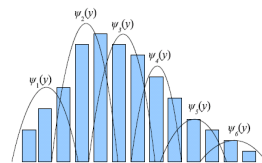
- IBM Researchの概要
- IBM Research数理科学部門の紹介
- IBM東京基礎研究所・数理科学部門の紹介
- プロジェクト事例: トラジェクトリ回帰とその応用
- まとめ

トラジェクトリ解析は、時空間データ解析、センサーデータ解析等で現れる応用上重要な問題

GPSTラッキングデータについての最近の研究例

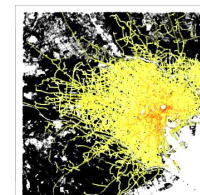


Map matching
[Raymond+ ICPR 12; Yanagisawa ICPR 10]

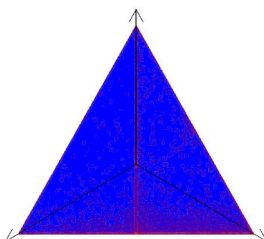


Travel time prediction
[Ide+ SDM 09; Takahashi SDM 12]

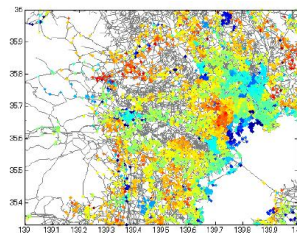
Location-based demand estimation
[Raymond+ AIGSPATIAL 11]



Charging station location optimization
[Yanagisawa IPDPS 10]



Human/driver personality estimation
[Osogami+ UAI 11, AAAI 12; Morimura+ ICML 10, UAI 10]



Origin-destination estimation
[Morimura+ ICPR 12]



Ultra-large scale traffic simulation for vehicles and humans
[Kato+ WSC 09; Mizuta WSC 09; Suzuki WSC 11]

Bus route optimization

ネットワーク上のトラジェクトリ回帰: ネットワーク上の任意の経路について「コスト」を予測するタスク

□ 入力: ネットワーク上の任意の経路

- 電子地図ならネットワーク上の隣接するリンクIDの系列として表せる

$$x^{(3)} = (25020201, 24021102, 222020101, 258020001, \dots)$$

link ID defined in digital maps

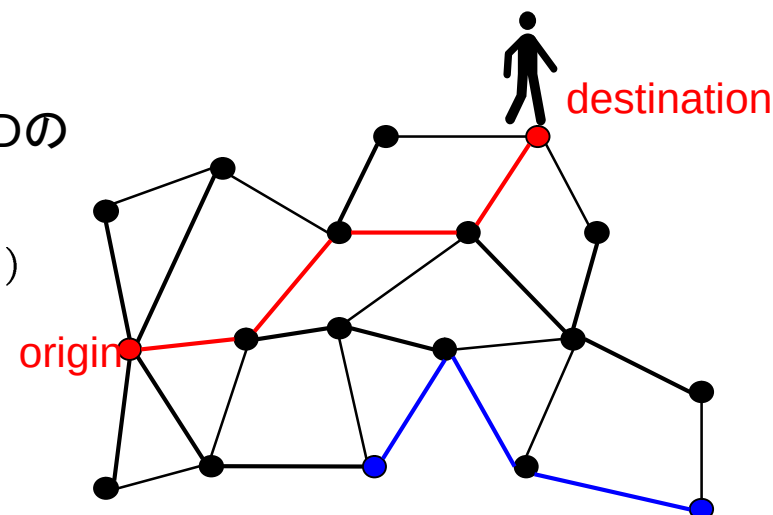
□ 出力: その経路のコスト

- スカラー値。所要時間などを想定。

□ 訓練データ:

$$\mathcal{D} \equiv \{(x^{(n)}, y^{(n)}) | n = 1, 2, \dots, N\}$$

- $x^{(n)}$: n-th trajectory (or path)
- $y^{(n)}$: n-th cost



regression function

$$y = f(\text{arbitrary path})$$

Cost

arbitrary path



ネットワークが明示的に定義されていない場合でも、このタスクに帰着できる問題は実用上いくつかある

□例1: 道路ネットワーク上の経路の旅行時間予測

- GPSではつねにとびとびの点でしか計測できないため、各リンクの所要時間を精度よく個別計測するのは不可能
- 経路全体 vs 所要時間、という回帰問題が自然

□例2: 蓄電池の劣化予測

- (後述)

□例3: 鉄鋼の圧延プロセスにおける消費電力予測



トラジェクトリ回帰についてのこれまでの研究例

□Idé & Kato, SDM 2009

- トラジェクトリ同士のカーネルを定義して
正規過程回帰の枠内で定式化

$$y = f(\text{trajectory})$$

□Idé & Sugiyama, AAI 2011

- Primalな側で、リッジ回帰として定式化
- (後述)

□Takahashi & Idé, ICPR 2012

- 蓄電池の劣化予測への応用

経路 x に含まれるすべてのリンクにわたる

The diagram shows a graph with 15 nodes and 20 edges. A red path of 5 edges connects the 'origin' node to the 'destination' node. A blue path of 4 edges connects a node at the bottom to a node at the bottom right.

$$c_e(f_e) = l_e(\textcolor{red}{f_e} + f_e^0)$$

リンク長(既知)

デフォルト値からのずれ
(未知)

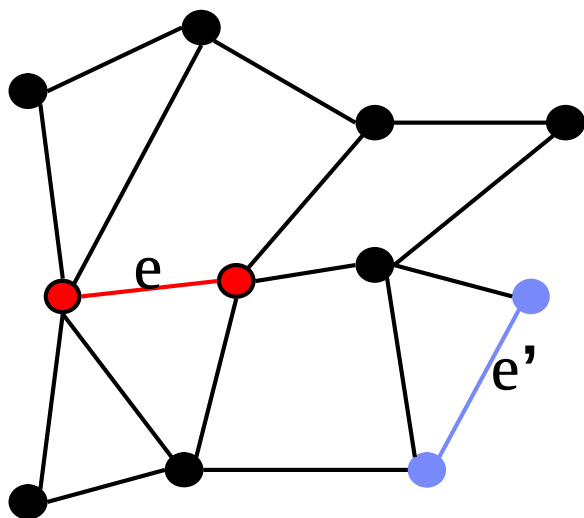
何が難しいか: 経路はほぼ無限の多様性があるので、すべての可能性を網羅することは不可能

□たとえば k最近傍法はうまく動かない

- 入力経路と非常に似た経路がないと、予測精度が上がらない
- 経路同士の類似度の定義にも任意性がある

□補足的情報を用いて情報の少なさを補う必要がある

□リンク同士の類似度を利用することを考える



$$S_{e,e'} \equiv \begin{cases} \omega^{1+d(e,e')}, & d(e,e') \leq d_0 \\ 0, & \text{otherwise} \end{cases}$$

d = (エッジ間のホップ数)

この場合 $d(e,e')=2$



似たリンクの特微量が似た値を取るような制約の下、トラジェクトリのコストの予測誤差を最小化する

□目的関数

$$\Psi(f|\lambda) = \sum_{n=1}^N \left(y^{(n)} - \sum_{e \in \mathbf{x}^{(n)}} c_e(f_e) \right)^2 + \frac{\lambda}{2} \sum_{e=1}^M \sum_{e'=1}^M S_{e,e'} |f_e - f_{e'}|^2$$

「予測したトラジェクトリのコストは、実測値と一致すべき」

- 「類似度が高いリンク同士では、特微量は似た値を取るべき」
- Sは、リンク間類似度をまとめた行列

目的関数を行列表現すると、ラプラシアン正則化項を持つ一種のリッジ回帰に帰着されることがわかる

$$\Psi(f|\lambda) = \left\| \mathbf{y}_N - \mathbf{Q}^\top \mathbf{f} \right\|^2 + \lambda \mathbf{f}^\top \mathbf{L} \mathbf{f}$$

$$\mathbf{y}_N \equiv [y^{(1)}, y^{(2)}, \dots, y^{(N)}]^\top$$

$$\mathbf{Q} \equiv [\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(N)}] \in \mathbb{R}^{M \times N}$$

類似度行列によるグラフ
ラプラシアン

$$\mathbf{L}_{i,j} \equiv \delta_{i,j} \sum_{k=1}^M S_{i,k} - S_{i,j}$$

$$q_e^{(n)} = \begin{cases} l_e, & \text{for } e \in \mathbf{x}^{(n)} \\ 0, & \text{otherwise} \end{cases}$$

$$[\mathbf{Q}\mathbf{Q}^\top + \lambda \mathbf{L}] \mathbf{f} = \mathbf{Q}\mathbf{y}_N \quad \text{これを解けばよい}$$

応用例1: 旅行時間予測 (Idé & Sugiyama, AAAI 2011)

実験の設定

□ ネットワークのデータ

- grid25x25 (25x25 square grid)
- 京都市街地図

□ 経路-コストデータ

- Nagel-Schreckenberg モデルという交通流のモデルに基づき交通シミュレーターで生成

□ 比較した手法

- Legal: 法定速度を完全を守る
- Simple: (略)
- GPR: 文字列カーネルを使った正規過程回帰
- Ours: 今回の手法



Figure 2: Downtown Kyoto City map of Kyoto data.

Table 1: Summary of data.

	Grid25x25	Kyoto
# nodes	625	1 518
# links	2 400	3 478
# generated paths	1 200	1 739

応用例1: 旅行時間予測 (Idé & Sugiyama, AAAI 2011)

シミュレーションデータによる比較

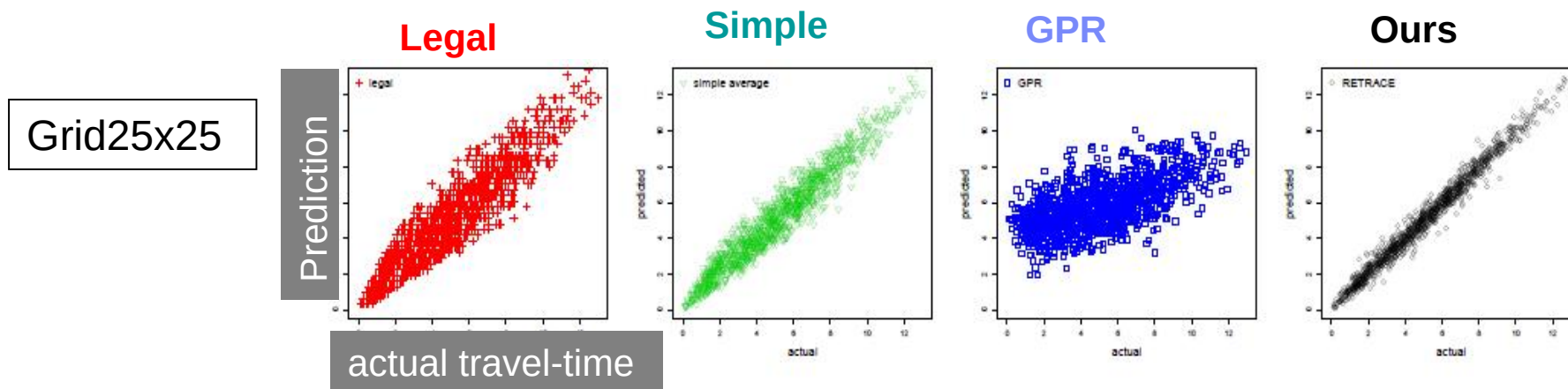


Figure 5. Comparison of travel-times in minute between Legal, Simple, GPR, and RETRACE (Grid25). The horizontal axis is the actual travel time, while the vertical axis is the predicted time.

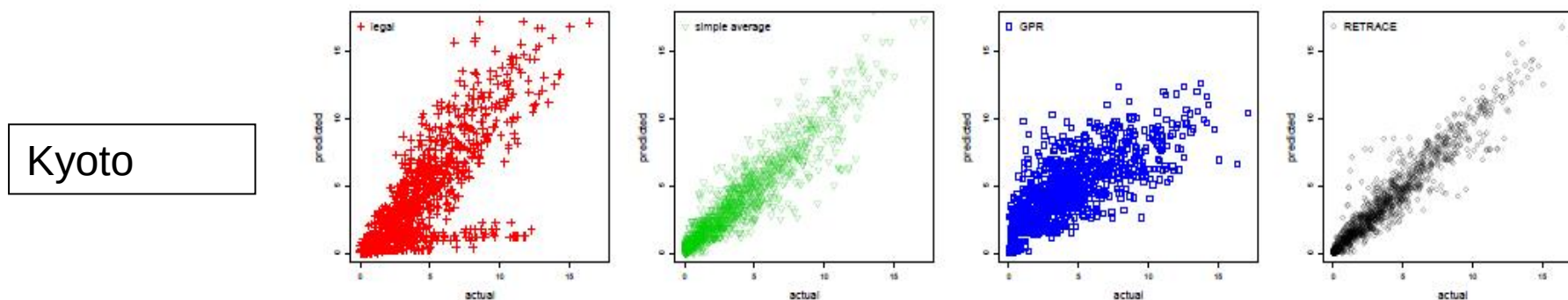


Figure 6. Comparison of travel-times in minute between Legal, Simple, GPR, and RETRACE (Kyoto). The horizontal axis is the actual travel time, while the vertical axis is the predicted time.

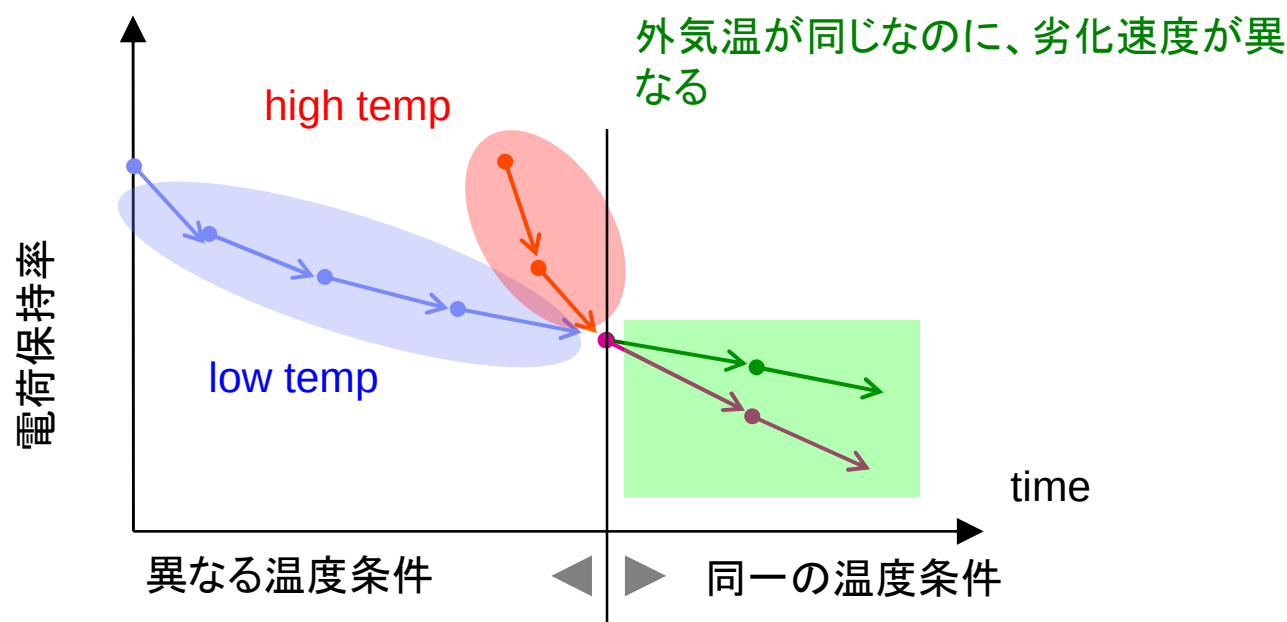
応用例2: 蓄電池の劣化解析 (Takahashi & Idé, ICPR 2012)

タスクは電荷保持率 (charge retention) の予測

□ 寿命予測における技術的チャレンジ

– 過去の使用履歴に寿命が強く依存する

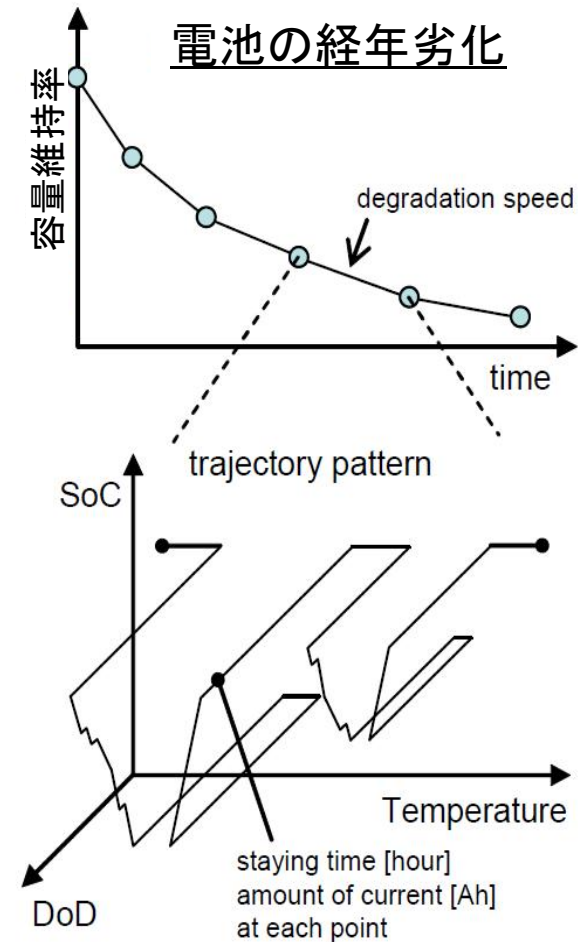
□ 例: 電荷保持率の温度依存性



応用例2: 蓄電池の劣化解析 (Takahashi & Idé, ICPR 2012)

蓄電池の使用履歴はトラジェクトリとして表現できる

- 電池の劣化度は、その時点で外部から観測可能な状態だけでは直接知ることができない
 - 電池の劣化度が過去の状態の履歴に依存するため
- 電池の使用履歴は、観測される物理量で張られる空間におけるトラジェクトリで表現される
 - 物理量: 負荷、温度、SoC (State-of-charge) が基本
 - SoC = 「フル充電に対し〇〇%」という量。ガソリン車の燃料ゲージと同様。



応用例2: 蓄電池の劣化解析(Takahashi & Idé, ICPR 2012)

トラジェクトリ回帰による電荷保持率の予測

□トラジェクトリ回帰により、過去の状態の履歴から、電池の劣化度を高精度に推定することが可能

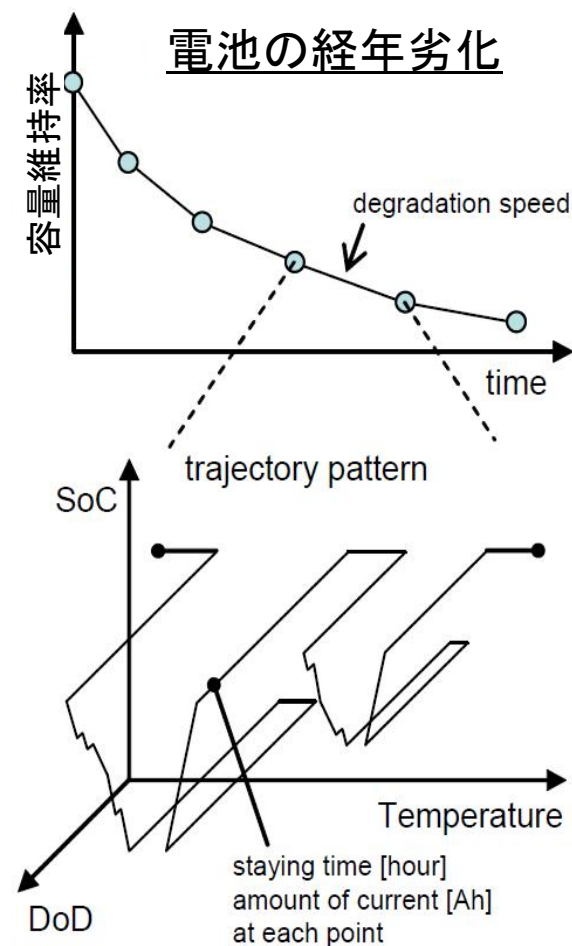
- 右下図: SoC、温度に加えて、「DoD」という特別な変数を採用した場合の、電池の使用履歴を表したもの
- 履歴を表すトラジェクトリ(軌跡)上の各点において、滞在時間と電流が記録されている

□過去に観測した(容量維持率, 物理量)の組から、トラジェクトリ回帰関数を構成することで予測が可能

$$y = f(\text{trajectory})$$

容量維持率

トラジェクトリ





応用例2: 蓄電池の劣化解析 (Takahashi & Ide, ICPR 2012) Li-ion電池における予測精度の検証結果

- 従来手法に比べて、予測誤差が数分の1に
 - 電池という成熟分野では顕著な改善



Contents

- IBM Researchの概要
- IBM Research数理科学部門の紹介
- IBM東京基礎研究所・数理科学部門の紹介
- プロジェクト事例: トラジェクトリ回帰とその応用
- まとめ



まとめ

- IBM Researchは、基礎研究と新規ビジネス開拓を両立させている
- 長い歴史と伝統を持つIBM Researchの数理科学部門は、あらゆる事業部門に主要な差別化技術を提供している
- IBM東京基礎研究所の数理科学部門は、機械学習・最適化・シミュレーションの専門家を揃え、その高い数理モデリング能力により重要なビジネス貢献を行っている
- センサーデータ解析のひとつの応用例として、トラジェクトリ回帰の蓄電池解析への応用事例がある