ビッグデータ応用
ブームから挑戦へ

JST CREST「ビッグデータ応用」研究総括
北海道大学大学院情報科学研究科
特任教授　　田中　譲

自己紹介

- 1974年京都大学電子工学専攻修士課程修了．同年北海道大学電気工学科助手．講師，助教授を経て1990年同教授．2004年同大学情報科学研究科教授，2013年同特任教授，名誉教授，現在に至る．1985年～1986年IBMワトソン研究所客員研究員．1996年より北海道大学知識メディアラボラトリ長．1998年～2000年京都大学情報学研究科併任教授．2004年より国立情報学研究所客員教授．2013年同運営委員会副会長，情報システム研究機構評議員．工学博士（東京大学）
- データベース理論（70年代），データベースマシン（70，80年代），知識メディアと知識連携（90年代以降）などの研究に従事
- 2006年よりEUのガン治験のIT統合支援システム研究開発プロジェクトに参加．2012年から文科省の目的解決型のIT 統合基盤技術研究開発プログラム「社会システム・サービス最適化のためのサイバーフィジカルIT統合基盤の研究」に参加し、札幌市の徐排雪の効率化を目指したビッグデータ分析を遂行．ビッグデータ応用技術、特に探索的可視化分析フレームワークの研究に従事．

ブームとしてのビッグデータ

- Volume: 大規模データに対応
- Variety : 多様なデータに対応
- Velocity : 実時間データストリームに対応
- Veracity : データの正確さが問題

- Value : 分析による価値創生

Volume： 大規模データ
Variety： 多様なデータ

- **センシング／モニタリング (sensor data)**
  - 農業（土壌・環境・生育状況）、漁業（漁場・漁獲）
  - 工業（生産ラインのモニタリング）
- **サービス利用ログ収集 (born digital)**
  - 都市基盤サービス： 電気・上下水道・ガス、公共交通
  - 情報通信サービス： ウェブ情報／サーチエンジン／モバイル通信／ナビゲーション
  - 金融サービス： クレジット・カード／ポイント・カード／保険／証券取引／投資ファンド
  - 医療サービス： 診断／治療／健康管理
- **観測・計測・検査データの取得・蓄積 (digital measurement and observation)**
  - 先端科学： 計測測定・観測・検査機器の自動化 → 大規模データ
- **大規模文献情報／ウェブ文書情報の蓄積**
  - 社会科学： ウェブ情報、SNS情報
  - 先端科学技術： 学術文献情報

Value:　分析による価値創生

- **センシング／モニタリング**
  - 1次産業:　環境センシング　→　収穫量最大化（ファクトリー化／最適画）
  - 2次産業計:　工程モニタリング　→　生産性向上
    - → **最適スケジューリング・最適制御**
    - → **状況変化と効果の予測**
- **サービス利用ログ収集**
  - 3次産業:　サービス提供　→　利用ログ収集　→　行動分析・意図抽出　→　価値創造
    - → **個人・集団・社会・市場の動向分析／意図抽出**
    - → **個人・集団・社会・市場の動向予測／リコメンデーション**
- **観測・計測・検査データの取得・蓄積**
  - 先端科学
    - → **摂理の発見**
    - → **現象の予測**
- **大規模文献情報／ウェブ文書情報の蓄積**
  - 社会科学
  - 先端科学技術
    - → **トピックス抽出と関係分析、トレンド分析、コミュニティ分析**
    - → **大規模知識ベースの構築　→　知識推論・知識発見**

Value：　分析による価値創生
これまでの成果

- 顧客・市場の動向分析・意図抽出
  - 製品販売／サービス提供
    - 仕入、販促、新製品開発への展開
    - 製造から販売までのラインの効率化
    - CRMの強化
  - 金融取引
- 生産工程の効率化
- 農業の収穫量向上
- 1次産業・2次産業と3次産業の直接連携によるトータルな効率化・収益増大
  - データを介した連携により、生産から流通、販売までのトータルな効率化・収益向上
- クレジット・カードの不正使用、不正侵入の検出

ビッグデータ応用の促進力
管理検索処理技術

- スケーラビリティ
  - Cloudを用いたHadoopなどの大規模データ基盤技術の発展
- 大規模データ検索処理
  - 列志向DBMSやNoSQLなどの新しいDBMS技術の進展
- 分析・可視化
  - 分析・マニング技術と可視化技術の急速な発展
- 大規模シミュレーション
  - 京コンピュータに代表される超高速計算技術と大規模シミュレーション技術の発展

ビッグデータ応用の促進力
機関によるデータの公開

- オープン化
  - 行政、公的機関、企業などのサイロの中に眠っていたビッグデータの積極的活用を目指したオープン化の動き
- 個人情報の保護と有効活用
  - 個人情報の取り扱いに関する法整備と解釈の基準化
    - 2014.5　**PCAST Report: Big Data and Privacy**

挑戦的ビッグデータ応用（1）

- 安心・安全でサステイナブルな都市基盤サービス
  - 交通／エネルギー／上下水道／ゴミ収集　etc.
  - Social cyber-physical system（SCPS)
    - IBM: Smater City / Microsoft: Urban Computing
  - 気象や交通、エネルギー消費などの過去データから規則的パターンを見つけ出し、これを実時間データに適用して予測を行い、効率的なサービス提供を目指す。
- 防災・減災・災害対策・復興支援
  - 地震／洪水／火災／テロ／…
  - 予測／予報／誘導／緊急対応支援／復興支援
  - 日常運用システムの機能拡張としての災害時緊急対応システム
  - 犠牲者・ボランティアを含む市民との連携支援機能
  - 臨機応変な対応、想定外の事態への対応が必要

挑戦的ビッグデータ応用(2)

- 感染症流行予測
  - 流行ウィルス株の予測
  - 流行の地理的拡散予測
- 農業の効率化・高収益化(e-agriculture) (→ 漁業・林業)
  - 生産・収穫支援
  - 流通販売支援
- ヘルスケア(特にバイオメディカル応用)
  - 個々人の病歴データの統合／病院や国の壁を超えたデータの統合管理分析
  - オーダメイド医療を目指した臨床治験・臨床試験データの分析
  - 創薬へのビッグデータ・アプローチ
- サイバー・セキュリティ

挑戦的ビッグデータ応用(3)

- 科学技術基盤システム
  - e-science / data intensive science / data centric science
  - 個別科学におけるビッグデータ応用

    宇宙物理学／地球環境学／高エネルギー物理学、核物理学／分子生物学、進化生物学／薬学／感染症疫学／材料物性学／…
  - 科学技術研究開発共通基盤システム技術
    - 多様なデータに基づく仮説設定／仮説検定の繰り返し過程の支援　←　分析法を見つけること自体が研究対象
    - 大規模な文献情報を推論可能な知識表現に直し、有益な知識を推論により発見（IBM　Watosnの機能を拡張発展）
      - IBM は、Watsonのバイオメディカル応用を重視

挑戦的ビッグデータ応用の特徴

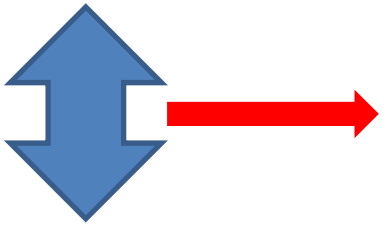
- Volumeに重点を置いた単一種類大規模データの分析

- Varietyに重点を置いた、多種類データを関連付けた分析による価値創生
  - 複数異種データを関連付けた分析
  - 異なるパターンや規則性に従う異種混合データの分析
  - 分析シナリオが定石として確立していない分野への応用
    - 先端科学技術分野や、戦略的意思決定を必要とする分野
    - 試行錯誤的・即興的分析過程の支援が必要

共通重要技術課題(1)

- 仮説設定／仮説検証の繰り返しの支援
  - 仮説設定: 分析対象を絞る（データ・セグメンテーション）
  - 仮説検定: 分析・マイニング・可視化
    → 対象集合を更にセグメンテーション

欠如！



大きなギャップ

スケーラブルな探索的可視化分析技術

- 多くの文献・書籍が示すビッグデータ分析過程:
  - データ収集 → データ整形・統合・表現 → データ分析・可視化・解釈 → 意思決定

共通重要技術課題（2）

- 大規模文献情報（コンテンツ情報＋リンク情報）
  → トピックス抽出と関係分析、トレンド分析、コミュニティ分析

- 大規模な文献情報を推論可能な知識表現に変換し大規模な知識ベースを構築することにより、有益な知識を推論により発見する技術
  - 1990年代に我が国の第5世代コンピュータプロジェクトが目指したことを、文献の取り込みから始め、大規模にスケールアップする試み
  - 専門家の支援だけでなく、映画「ロレンツォのオイル」の両親のように、難病の治療法を公刊されている膨大な学術論文を渉猟して発見しようとする素人をも支援できるような技術の開発
  - 現在の技術でこれにもっとも近い位置にいるのがIBM社のWatson

私自身のビッグデータ応用プロジェクト

- ガンの臨床治験の統合IT支援
  - EUのFP（フレームワーク・プロジェクト）
    - FP6 Integrated Project ACGT (Advancing Clinico-Genomic Trials on Cancer) (02/2006 – 07/2010)
      - 26 teams
    - FP7 Large-scale Integration Project p-medicine (personalized medicine) (02/2011 – 01/2015)
      - 29 teams
- 都市環境モニタリングと社会サービスの管理
  - 文科省：ソーシャルCPSプロジェクト (09/2012-03/2017)
    - NII、北大、阪大、九大
    - 札幌市の徐排雪の効率化・最適化

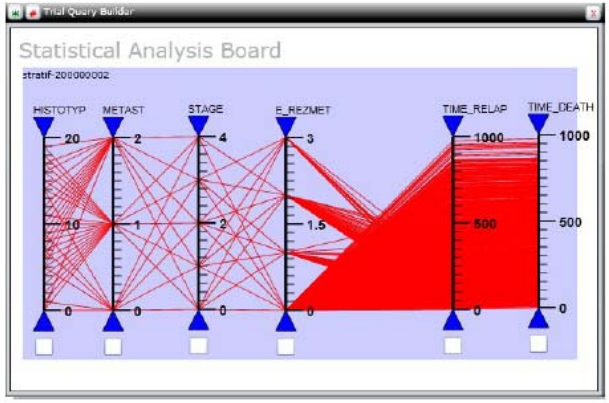
Trial Outline Builder (TOB) (2010)

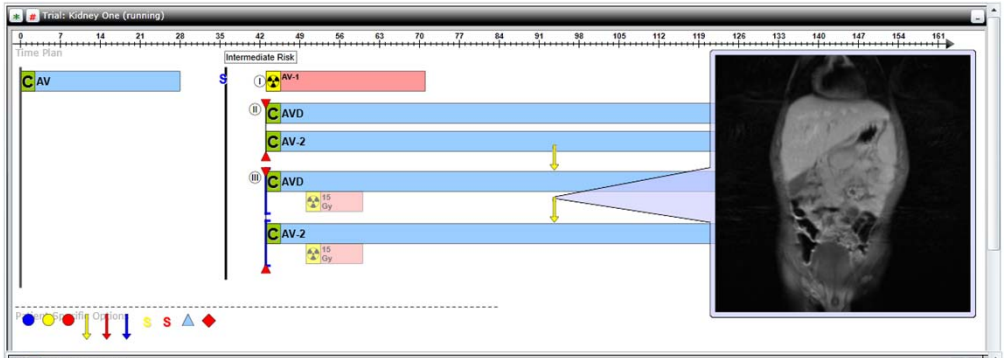
(EU 第6期フレームワークプログラム：大規模統合プロジェクトACGT (患者遺伝子データも用いたガン臨床治験の統合支援システム)

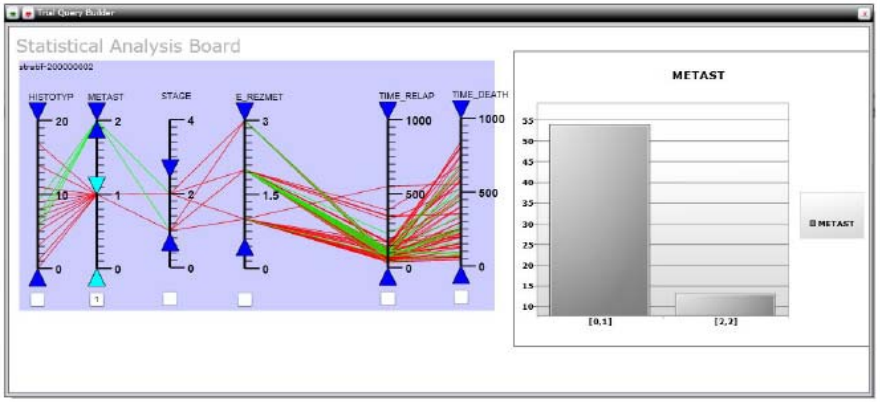
- 臨床治験プラン編集環境
  - 作図システムを用いるように、治療イベントや診断イベントをドラッグ・アンド・ドロップしてフローを定義
  - 各イベントをクリックしてカルテを定義
  - 定義結果よりデータベースをシステムが自動定義
- 患者治療環境
  - 個々の患者の治療・診断をガイド
  - 各イベントの患者データ入力をガイド
  - ランダマイゼーション後の候補治療はシステムが患者ごとにランダムに選択
- 検索・分析環境
  - 平衡座標系や相互に連携した複数のビューを用いて、探索的な可視化分析過程の遂行を支援

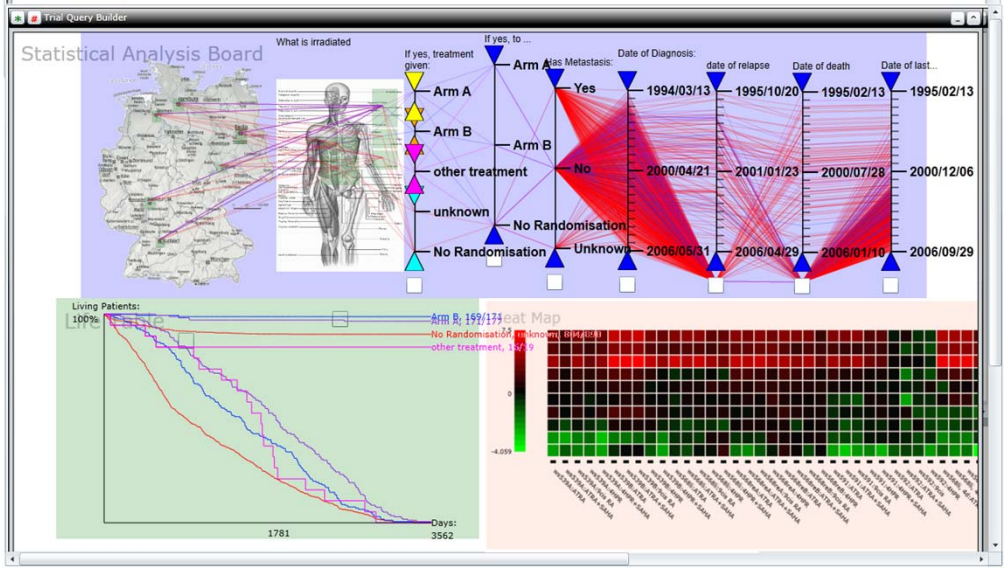
探索的可視化分析

Trial Outline Builder (TOB) : Query & Analysis View



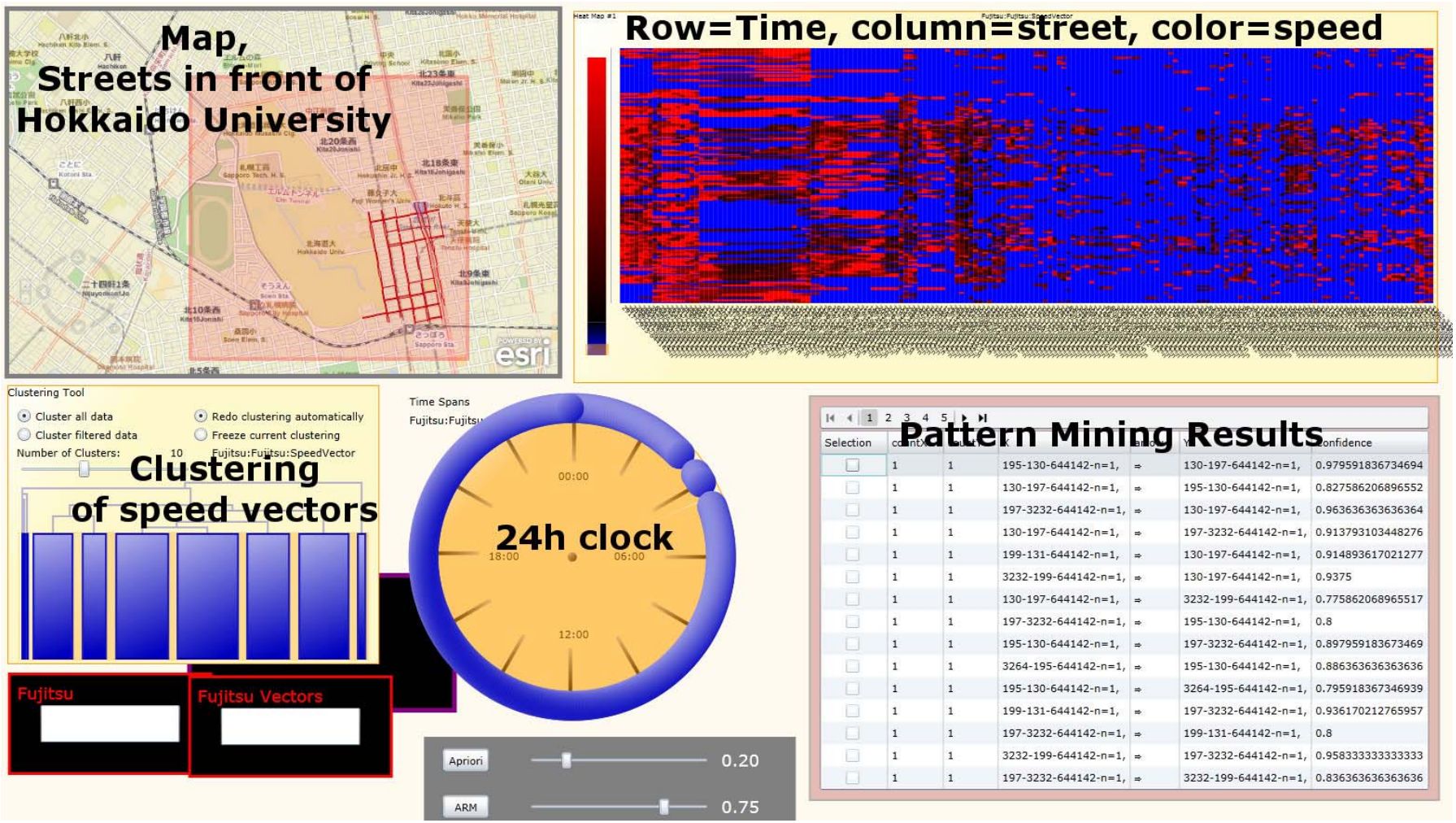






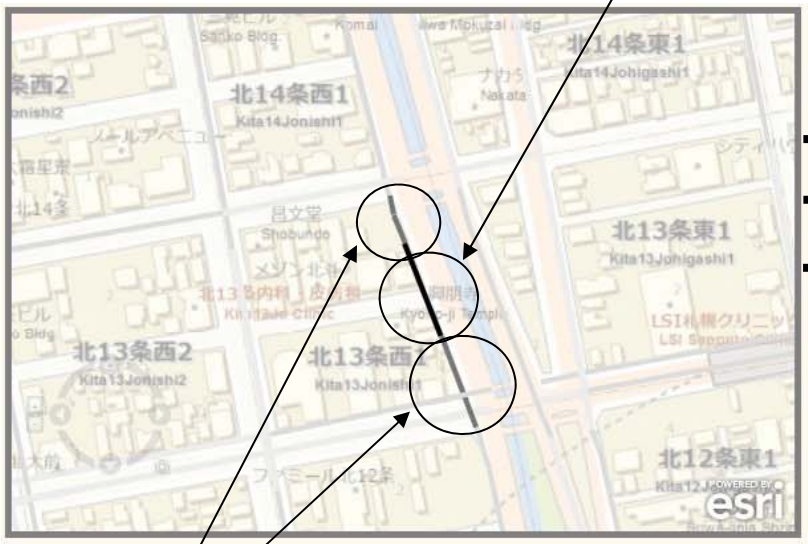
探索的可視化分析
Geospatial Digital Dashboard

5分ごとの道路リンクごとのタクシーの平均速度データを分析



マイニングの結果得られた渋滞の依存関係

前件となる渋滞個所



Selection	countX	countY	X	arrow	Y	Confidence
☐	1	1	195-130-644142-n=1,	⇒	130-197-644142-n=1,	0.933333333333333
☑	1	1	130-197-644142-n=1,	⇒	195-130-644142-n=1,	0.823529411764706
☐	1	1	197-3232-644142-n=1,	⇒	130-197-644142-n=1,	0.977777777777778
☑	1	1	130-197-644142-n=1,	⇒	197-3232-644142-n=1,	0.862745098039216
☐	1	1	3232-199-644142-n=1,	⇒	130-197-644142-n=1,	0.866666666666667
☑	1	1	130-197-644142-n=1,	⇒	3232-199-644142-n=1,	0.764705882352941
☐	1	1	3232-199-644142-n=1,	⇒	197-3232-644142-n=1,	0.888888888888889
☐	1	1	197-3232-644142-n=1,	⇒	3232-199-644142-n=1,	0.888888888888889
☐	1	1	3232-199-644142-n=1,	⇒	199-131-644142-n=1,	0.955555555555556
☐	1	1	199-131-644142-n=1,	⇒	3232-199-644142-n=1,	0.977272727272727
☐	1	1	3193-202-644142-n=1,	⇒	202-3226-644142-n=1,	0.759259259259259
☐	1	1	202-3226-644142-n=1,	⇒	3193-202-644142-n=1,	0.745454545454545
☐	1	1	3193-202-644142-n=1,	⇒	204-3193-644142-n=1,	0.740740740740741
☐	1	1	204-3193-644142-n=1,	⇒	3193-202-644142-n=1,	0.869565217391304
☐	1	1	2954-2972-644142-n=1,	⇒	2972-8966-644142-n=1,	0.738461538461539

後件となる渋滞個所

平成25年度採択課題と平成26年度の重点領域

- 平成25年度採択課題
  - 医薬品創薬から製造までのビッグデータからの知識創出基盤の確立
    - 船津　公人　(東京大学 大学院工学系研究科　教授)
  - 「ビッグデータ同化」の技術革新の創出によるゲリラ豪雨予測の実証
    - 三好　建正　(理化学研究所 計算科学研究機構　チームリーダー)
  - 両方ともビッグデータの一部は計算機により生成
  - モデリングによるシミュレーションと実測データの同化 → 予測
- 平成26年度重点領域
  - オーダメイド医療（personalized medicine)
  - 防災・減災・災害対応
  - スケーラブルな探索的可視化分析
- 平成27年度
  - 農業・漁業（生産・流通・販売）へのビッグデータ・アプローチ
  - 先端科学技術への応用
    - 個別分野における応用
    - 共通基盤技術の高度化
      - 探索的可視化分析環境技術
      - 大規模な文献情報を推論可能な知識表現に変換し大規模な知識ベースを構築することにより、有益な知識を推論により発見する技術