プロジェクト名：　情報化時代にめざす科学的推論の形［機能と帰納］

プロジェクトディレクター：　樋口知之

１．研究目標

情報社会の実現によって様々な分野で複雑なシステムに関する大量データに基づく予測と，そのリスク評価の方法の確立が重要な社会的課題となっている。本プロジェクトでは，地球，生命，社会等の４研究所の融合分野において戦略的研究を推進しながら，複雑なシステムの理解のための，帰納的手法，あるいは帰納的手法と演繹的手法との融合的手法による，システムの機能のモデル化に関する研究開発を行う。ここで機能のモデル化とは，対象そのものを実体的に精緻にモデル化するのでなく，対象に関する情報の入出力関係に代表されるような，機能自体を模倣する数理モデルを構築することを意味する。これにより，統計的モデル構築法と予測アルゴリズム，情報抽出・知識発見のための情報統合の方法など，分野に共通の道具を生み出すことを研究の目的とする。

またこれらのモデリングの方法を基盤として，リスクの（科学的）評価と管理のための方法論の確立をすすめる。そして，情報とシステムという視点から不確実性に関わる研究の新分野を開拓し，現代社会が直面する諸問題の解決を通じて，安心・安全な国家社会の構築，地球環境の改善など持続的な繁栄を目指す。

２．研究概要

本プロジェクトでは，地球，生命，社会等の４研究所の融合分野において，統計的モデル構築法と予測アルゴリズム，情報抽出・知識発見のための情報統合の方法など，分野に共通のツールを生み出すことを目標としている。

２．１　*Statistical Thinking Machine* をつくる

２．１．１　言葉の整理

あらゆる分野において，大量データからいかにして知識発見を自動的にかつ効率よく行うかが研究推進の鍵となってきている。超大規模データの取り組みにあたって，この目的達成のために必要とされる作業（プロセス）の体系化が本プロジェクトの大きなねらいの一つである。体系化に際しては，言葉の整理をしてみるのも研究進展の大切な一歩である。樋口(PD)は統計科学，情報学関連の研究者と意見交換をする中で，各研究分野において同じ言葉を違った意味で用いていることに気づいた。図１はそれらを図化したものである。上の階層ほど集約度が高い普遍的な価値をもつ量である。横の階層は同じ概念で，左が統計科学関係者が使う言葉，右が情報科学関係者が使う言葉である。この図からわかるように，統計科学関係者が「データ」とよぶものは情報科学関係者では「情報」に，また「情報」と呼ぶものは「知識」に対応する。そうすると，統計科学関係者が「知識」と呼ぶものに対応する，情報科学関係者一般で使用されている言葉はない。あえて言えば，各応用分野における「○○知」といったものであろうか。「英知(Wisdom)」と英

語で呼ばれることもある。方や，情報科学関係者がデータと呼ぶ，センシングからの直接的な生データに対応する言葉は実は統計科学にはなく，それは統計科学関係者があまり取り扱ってこなかったことを裏付ける。

本プロジェクトの目的は，下の階層から上の階層のものを導く手法（図中では紫色の矢印で表記した）の開発や方法論の研究である。これがまさに分野共通のツールとなりえる。将来的には，センシングからトップレベルまでを直接的につなぐ方法の開発を念頭においているのは言うまでもない。

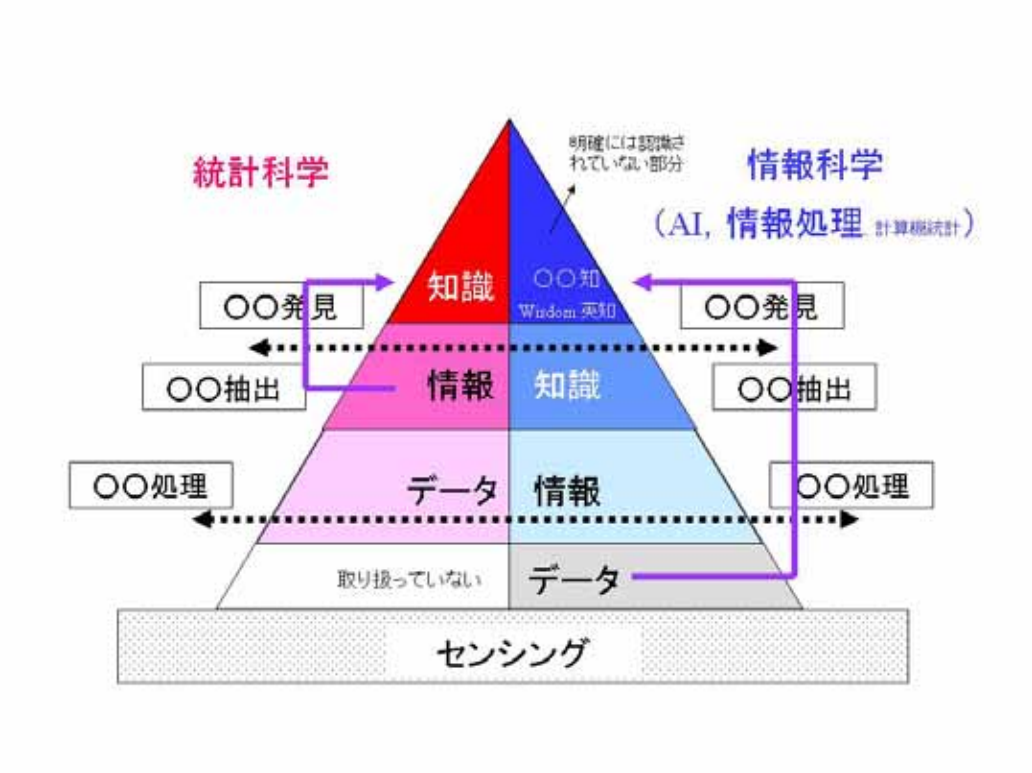

図１：言葉の使われ方

２．１．２　そのほかの研究プロジェクトの現状

科学研究費特定領域研究「情報爆発」（http://www.infoplosion.nii.ac.jp/info-plosion/ ）における研究では，テキスト，（動）画像といった，Web 空間に特徴的な対象に焦点をあわせがちである。これからは，これまでの情報科学で取り扱いやすかったそのような対象に関する研究から離脱し，自然科学全般を取り扱うのは当然で，さらに日常生活まるごとを対象にした研究推進の端緒に着くべきであろう。それは，NSF の CDI（Cyber-enabled Discovery Innovation http://www.nsf.gov/crssprgm/cdi/）や，QoLT(Quality of Life Technology　http://www.qolt.org/)の研究開発理念をあわせ，さらに拡大発展させたようなプロジェクトであることが望ましい。

CDI のプロジェクトでは，”Computational Thinking” に基づく技術革新や発展により革新的な科学・工学の研究成果物を生み出すことを目的としている。この”Computational Thinking”を NSF は以下のように定義している。

Computational thinking is defined comprehensively to encompass computational concepts, methods, models, algorithms, and tools.

これはまさに我々のプロジェクトが狙う分野共通の道具と重なるものである。我々のプロジェクトはこの CDI の概念と重なる部分も多いが，“機能と帰納”という統計的概念を強調している点が独自的である。それゆえ，我々のプロジェクトのねらいは，”Statistical Thinking Machine”をつくることとも言える。CDI では研究開発する３つの大きな研究テーマをかかげている。

- **From Data to Knowledge:** enhancing human cognition and generating new knowledge from a wealth of heterogeneous digital data;
- **Understanding Complexity in Natural, Built, and Social Systems:** deriving fundamental insights on systems comprising multiple interacting elements; and
- **Building Virtual Organizations:** enhancing discovery and innovation by bringing people and resources together across institutional, geographical and cultural boundaries.

一番目が，２．１．１の２段落で説明した目的と一致している。

２．２　分野共通のツールの開発

研究代表者（プロジェクトディレクター）のリーダーシップのもとにプロジェクトを機動的に推進するため，３つのサブ研究テーマを策定している。各サブテーマ名と，各々が関連する手法---分野に共通，あるいは他分野へ転用が平易なツール---を以下に示す。

－ 具体的なサブプロジェクト －－

Ａ．予測とリスク解析

関連する手法：データマイニング，テキストマイニング，Web マイニング，グラフモデリング，機械学習，ゲノム配列解析，バイオインフォマティクス，DNA アレイデータ解析手法，金融・経済システムの研究，ファイナンス数理，保険数理，医薬品・食品安全性の研究，疫学，サンプリング調査法，環境リスク研究，災害リスク研究，極値分布の研究，セキュリティに関する研究，電子商取引の制度

Ｂ．情報・通信“メタウェア”とその応用

関連する手法：グリッド計算技術，乱数に関する総合研究（物理乱数，擬似乱数），並列計算技術，統計ソフトウェアの共通基盤化，諸ゲノム解析手法の R への実装，数値的最適化技法，音声・画像解析，音声・話者認識，自然言語処理，ロボティクス，対話技術，デジタル・アナログ信号処理技術の融合，マルチパス干渉，マルチユーザ干渉，超高速データのコヒーレント伝送技術

C．ダイナミック逆問題

関連する手法：オーロラ画像解析，映像処理，画像合成，ベイジアンモデリング，オーロラ３次元構造復元，地球環境長期予測，巨大次元の数値解法，時系列解析手法，画像解析手法，時空間モデリング，データ同化，レーダー観測信号処理，諸ノイズ除去手法，異常値処理，時空間モデリング，三次元形状のモデル化，高解像度シミュレーション

２．３　サブプロジェクト構成

表１にサブプロジェクトの構成上の要点をまとめた。R，G，Bの三色を混ぜると白色光になるように，３つのサブテーマを束ねるとどの分野にも共通につかえる道具，いわば無色透明的存在物を生み出す研究プロジェクトであることを示している。

	機能と帰納プロジェクト（↓のRGBを混ぜると、無色に）		
	G: Green	R: Red	B: Blue
	サブテーマA	サブテーマB	サブテーマC
サブプロジェクト(SP)名	予測とリスク解析	情報・通信"メタウェア"とその応用	ダイナミック逆問題
サブプロジェクトディレクター(SPD)	江口教授(統数研)	中野教授(統数研)	佐藤　副所長(極地研)
機構内で中心となる研究所	(統数研，遺伝研)	(統数研，情報研)	(極地研，統数研)
中心的(応用)領域	生命・環境科学	情報工学	地球科学
中心的最終成果物(*1)	データベース(*2)	ハードウェア，ソフトウェア	新しい観測データ
(*1): プロジェクト全体(サブテーマ共通)の成果物以外の，サブテーマに特徴的な成果物をさす．			
(*2): リスク解析センターの医薬品・食品リスク研究グループは，医薬品の安全性に関するデータベースを作成予定．また金融・保険リスク研究グループもデータベースに関連した研究成果を出す予定．			

表１．　サブプロジェクトの特徴

３．年度計画

テーマ	１６年度 予備研究	１７年度 プロジェクト初年度	１８年度 研究レビュー	１９年度 中間評価	２０年度	２１年度
研究体制の編成	←	―	→			
情報収集・整備		←	―	→		
研究会ワークショップの開催		←	―	―	―	→
研究体制の見直し			←	―	→	

平成１６年度（予備研究）

平成１６年度に新領域融合研究センター研究課題として採択された以下の研究プロジェクト

を，１７年度スタートした本プロジェクトは引き継いでいる。

・オーロラ科学における画像解析と逆問題

研究代表：　佐藤夏雄（国立極地研究所）

・帰納機械による動的なマルチモーダル情報の検索と認知の研究

研究代表：　松井知子（統計数理研究所）

・南極大型大気レーダーによる高級観測アルゴリズムと高速データ処理システムの開発

研究代表：　江尻全機（国立極地研究所）

・磁気圏・電離圏・大気圏複合システムの定量的解析に向けた研究

研究代表：　樋口知之（統計数理研究所）

・科学研究における計算機によるモデリング環境

研究代表：　中野純司（統計数理研究所）

平成１７年度(プロジェクト開始)

統計的モデル構築法と予測アルゴリズム等分野に共通の道具を生み出すために，まずサブテーマを選定し集約的に効率よく研究をすすめる体制を整える。特に，予測と発見のためのモデリング技術とアルゴリズム開発，計算機による帰納的モデリングのための環境開発，マルチモーダルデータからの不変情報の発見とその方法，高速データシステムのモデル化技術，リスク解析とその評価技術など分野横断的な研究の情報収集，整理を網羅的に行う。

平成１８年度

スーパーコンピュータ上での並列化と（物理及び疑似）乱数の利用が比較的簡単な操作で行えるようなシステムの開発や，高速データ通信モデル化用の準備的なハードウェアを試作する。マルチモーダルデータに含まれる不変情報の発見や,大規模アレイデータからの構造抽出に関する，帰納的手法を用いた既存手法の整理をする。個別科学におけるリスク解析の現状とのギャップをはかるため，各分野で比較的小規模のワークショップを複数開催し分野間連携の準備を行う。

平成１９年度(中間評価)

データのコーディングモデル,アルゴリズムやモデルなどについて,分野横断的な考察を行う。また,分野間の交流と統計解析手法の水準のボトムアップを引き続いて図る。スーパーコンピュータ上での並列化と（物理及び疑似）乱数の利用が比較的簡単な操作で行えるようなシステムの開発を引き続いて行う。システムをRに限定せずにデータの可視化や解析を行うための研究も行う。マルチモーダルデータに含まれる不変情報の発見や，大規模アレイデータからの構造抽出に関する，帰納的手法を用いた既存手法の発展を図る。また，マルチパス，フェーディング環境下で高速・高性能を実現するための無線システムモデル化の研究を推進する。アレイ観測データの効率的なノイズリダクション法や，地球科学データのダイナミック逆問題解法の研究をすすめる。

平成２０年度

複数のスーパーコンピュータやパーソナルコンピュータが有機的に協力してモデリングを行えるような環境の研究を行う。帰納的メタウェア，データのコーディングモデルに関する考察結果の整理や,開発した観測アルゴリズムによる試験観測などを通して,帰納的手法の体系化を行う。リスク解析に関わる外国人客員の招聘等により，チュートリアルセミナーを開催することによって，研究者養成に資する。

平成２１年度

機能と帰納での科学研究におけるモデリングを行う際の有用なツールとなるシステムの例示や，帰納的メタウェアやマシンのツール化，資料化を行う。開発した帰納的手法のインターネット等を通した一般公開や，高速・高品質無線伝送システムモデルのフィールド評価を行う。またあわせてこれらの手法の他分野への適用について研究する。成果は適宜国内外の学会，及び論文にて発表する。高速・高品質無線伝送システムのモデルを確立し，国内外へ発表・提案・啓蒙活動を行う。アレイデータ観測実システムへの適用を国内外に対して提案し，成果発表・チュートリアル等を行っていく。過去４年間の研究成果を踏まえ，必要に応じて既存研究サブテーマの見直しと人員の再配置を行い，新規重点分野の開拓を行う。複雑なシステムの理解を加速する，アルゴリズムとモデリング技術のさらなる研究開発の推進により，安心・安全な国家社会の構築，地球環境の改善など国家及び人類の持続的な繁栄に貢献していきたい。

平成２２年度以降の展開

●サイエンティフィック（学術）・サービス・イノベーションセンター（SSI）の設立へ

①　中核的機関へ

中期計画第二期では融合センターの守備範囲を経済・社会分野にまで拡張し，統数研のこれまで蓄積した幅広い研究分野における共同研究のノウハウを生かして融合センターを運営する形態も考えられる。複合，融合という言葉がつく，分野を超えた連携が本質的に必要な学問領域の創造，育成，発展には，まずこのセンターの人的ネットワークが活用されるよう目論む。いわば，サイエンティフィック（学術）・サービス・イノベーションセンターを立ち上げる。

②　育成：　大規模データ活用人材育成

このセンターの目的に沿って本格的に活動できる人材は，日本は残念ながら少ないと言わざるを得ない。その理由は，昔から言われている縦割り的研究分野や学会のあり方や，演繹的思考をたたきこむ大学教育現場の影響である。書物，講義を通してだけのデータ解析の習得では現場で応用のきく実力（地力）はつかない。共同利用や協力をこえる形態，分かりやすく言えば次世代の若者をシゴク“現代版道場”をセンター内に開き，新しい環境，問題に対して柔軟に立ち向かい，かつしぶとくねばることのできる人材を養成する。別の人材タイプとして，異分野の研究が複数わかり共同研究のコーディネータができる，いわば科学のマルチリンガルを育成する。このタイプの人材は，言わば“超常識人”，あるいは“超雑学人間”といえよう。

ただし，昔の教養学部の教育方針と誤解されぬよう十分な注意が必要である。

③　新分野開拓

上の①，②を束ねることで新しい新分野開拓に対応する。

４．研究費の推移

平成１７年度実績：　148,080 千円

平成１８年度実績：　124,663 千円

平成１９年度見込：　147,890 千円

５．平成１９年度の研究推進体制

（1）予測とリスク解析

研究代表者

［統計数理研究所］　江口真透

共同研究者

［大阪大学・産業科学研究所］　鷲尾　隆

［統計数理研究所］　江口真透　　足立　淳　　椿　広計

（2）情報・通信“メタウェア”とその応用

研究代表者

［統計数理研究所］　中野純司

共同研究者

［統計数理研究所］　中野純司　　松井知子　　瀧澤由美

（3）ダイナミック逆問題

研究代表者

［国立極地研究所］　佐藤夏雄

共同研究者

［国立極地研究所］　門倉　昭　　和田　誠

［統計数理研究所］　尾形良彦

６．平成１９年度の研究進捗

プロジェクトがスタートして実質的に２年経過し，また平成１９年度始めに行ったサブテーマの組み替えも予算管理運営上と研究体制上ともにうまく機能したことで，研究成果も国際会議録を始めとしてとして少しずつ出てきてきた。成果報告の情報発信として重要なホームページも，このサブテーマ構成の変更にともない，すみやかにその情報修正をするとともに，すべてのサブサブテーマ的な研究内容レベルにいたるまで詳細な研究紹介のホームページを公開することがで

きた。各サブテーマ主催の研究会やワークショップなどをとあわせて，ホームページの充実により，各サブテーマの具体的な成果が多くの方々に理解してもらえたと考える。

プロジェクト内の予算執行に関する管理運営については，機構の融合センター事務等と綿密に情報交換を行いながら毎年のようにその改善に努め，平成１９年度は特段の改良を行った。具体的には，全体的予算総額を年度当初にサブテーマに配分するのでなく，PD 預かり分を相当分確保しておき，年３回ほど定期的に各サブテーマに予算申請を行ってもらうことで，弾力的かつ柔軟な予算執行が可能となる体制とした。また，平成１９年度秋の融合研究シンポジウムにあわせて，新サブテーマ構成のもとでの研究体制の説明と，これまでの成果紹介のためのカラーパンフレットを作成した。同時に，本プロジェクトの全体研究目的に，各サブテーマがどのような観点から貢献しているのか，あるいは大学共同利用機関における本プロジェクトの位置づけと役割等を分かりやすく解説した資料も作成し，多数の方に配布した。

７．平成１９年度の研究成果

（1）成果物（知見・成果物・知的財産権等）

以下，主要成果のみ列挙する。詳細はサブテーマ欄を参照していただきたい。

- 機械学習の方法論を発展させ応用分野の研究者と協力することで，初期段階のガンの検出に有効であるプロテオームデータ解析システムを提案（予測とリスク解析）
- 市販後医薬品の有効性･安全性の科学的評価のため，大規模なデータベースを日本で初めて構築し，いくつかの活用例を示すことでデータベースの有用性を実証（予測とリスク解析）
- 統計科学でデータ解析の標準的なシステムとなっている統計解析システムＲを，本機構の保有するスーパーコンピュータで並列利用可能にすることで計算効率性を高め，プログラム等の成果物をＲのコミュニティに還元（情報・通信“メタウェア”とその応用）
- 映像検索等の具体的なテーマに対して統計的機械学習の有用な手法を開発し，国際的な評価を獲得（情報・通信“メタウェア”とその応用）
- 統計科学と情報科学の研究者が無線通信ハードの開発者と共同研究体制を組むことで，統計手法を組み込んだ複雑な状況下での高速データ通信システムのプロトタイプの開発に成功（情報・通信“メタウェア”とその応用）
- 地球の磁力線が南北半球間で繋がっている地点（地磁気共役点）は時々刻々変化することが理論/モデルで予測されていたが，可視オーロラの南北同時観測により，世界で初めて，共役点位置の時間変化を正確に観測事実から証明（ダイナミック逆問題）
- 昭和基地とアイスランドのチョルネス観測点で同時に観測された多種多様な脈動オーロラの統計的データ解析を行い，脈動オーロラについてはその非共役性を実証(ダイナミック逆問題)
- 汚染大気の大陸間輸送等を調べる粒跡線モデルをオンラインで利用可能なシステムを web 上に構築（ダイナミック逆問題）

・長期間にわたり大量に収録されている気象庁などの震源カタログを包括的に取り扱える，究極の地震活動解析用の時空間点過程モデルを構築（ダイナミック逆問題）

（2）成果発表等

＜会議発表等＞

〔招待講演〕

・2007 年 9 月 6 日

樋口知之，「大規模データ解析の現状と問題点」統計関連学会連合大会

・2007 年 11 月 6 日

樋口知之，「統計モデルを用いた大規模データの分類，変換，そして知識発見」，第 10 回 情報論的学習理論ワークショップ（IBIS 2007）

〔一般講演〕

・2008 年 2 月 27 日

樋口知之，『機能と帰納プロジェクト』の紹介，統計数理研究所共同利用研究重点テーマ「統計メタウェアの開発」共通公開研究会

＜著書等＞

・樋口知之 監修&著 他著，ベイジアンモデリングによる実世界イノベーション『統計数理は隠された未来をあらわにする』，東京電機大学出版局（2,200 円税別），2007。

・樋口知之 他 翻訳&監訳，パターン認識と機械学習 上 - ベイズ理論による統計的予測，シュプリンガー・ジャパン（6,500 円税別），2007。

＜受　賞＞

該当無し

（3）その他の成果発表

・2007 年 6 月 29 日，科学新聞（2007 年 6 月 29 日号）の書評欄に，『統計数理は隠された 未来をあらわにする』がとりあげられる。

・2007 年 11 月 24 日，朝日新聞土曜版 Be（ビジネス Be Report）にベイジアンモデリングの特集が掲載。取材に協力。談話も掲載される。