BBL@RIETI
平成24年5月14日(月)

ビッグデータが迫る研究開発の変革

樋口知之　（情報・システム研究機構　統計数理研究所）

統計数理研究所の概要

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

設置目的・沿革・活動

設置目的

○ 統計数理に関する総合研究

沿　革

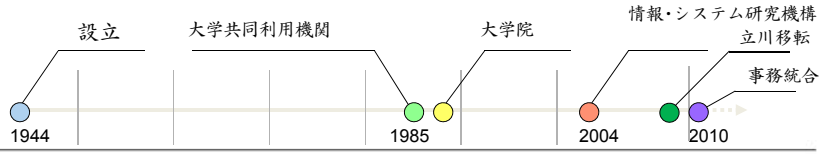
- ●1944年：文部省直轄の研究所として設立
- ●1985年：大学共同利用機関に改組転換
- ●1988年：総合研究大学院大学創設
- ●2004年：法人化，機構化
- ●2009年：立川移転
- ●2010年　極地研と事務統合。Akaike Guest House。

主要活動

○ 研究活動
- ・我が国における統計数理の中核拠点
- ・先端的な研究を推進

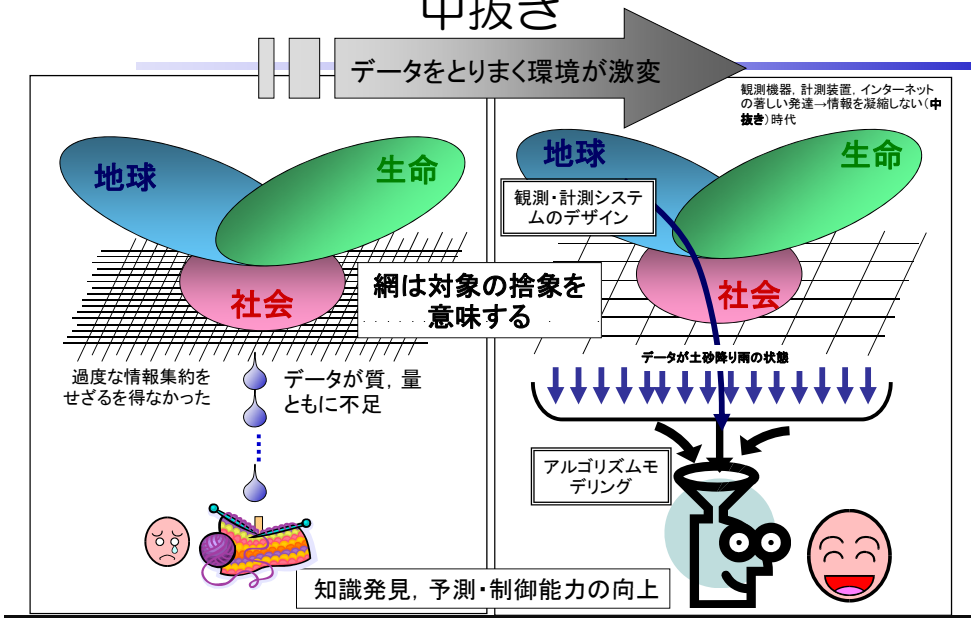
○ 共同利用
- ・多様な分野との共同研究
- ・スパコン，ソフトウェア，乱数

○ 人材養成
- ・総研大における大学院教育
- ・公開講座



大学共同利用機関法人 情報・システム研究機構 統計数理研究所

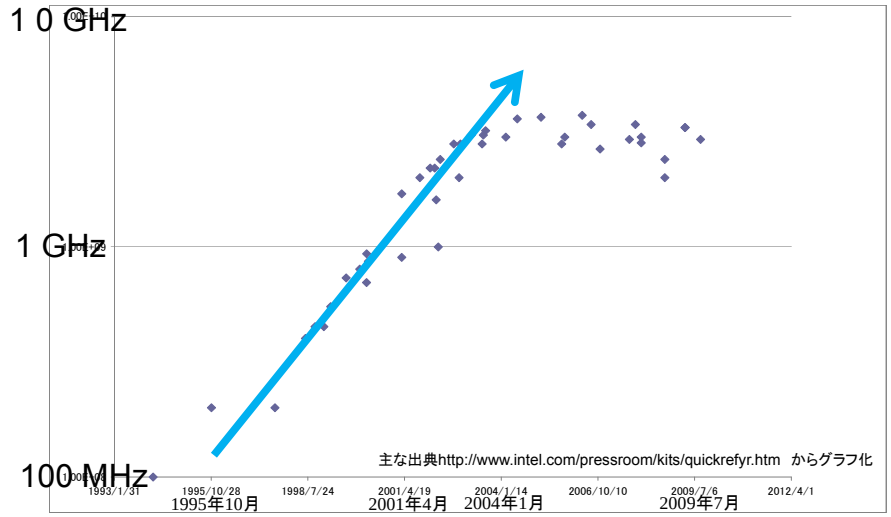
中抜き



大学共同利用機関法人 情報・システム研究機構 統計数理研究所

演繹推論をささえる計算機の性能の動向は？

単体CPU（コア）のクロック数の限界



※2010/Feb IBM Power 7 4.1 GHz

人生をハードディスクに埋め込む

10分ごとに1枚写真をとると、
5MB×6×24×365×80 ≅ 20テラバイト

2テラバイト　10,980円

11万円で一人のメモリーが記録可能

ビッグデータとは？

Researchers in a growing number of fields are generating extremely large and complicated data sets, commonly referred to as "big data."

http://www.nsf.gov/news/news_images.jsp?cntn_id=123607

課題：　気象学、ゲノミクス、コネクトミクス、複雑な物理シミュレーション、環境生物学、インターネット検索、経済学、経営情報学

データの源：　モバイル機器に搭載されたセンサー、リモートセンシング技術、ソフトウェアのログ、カメラ、マイクロフォン、RFIDリーダー、無線センサーネットワーク

ウィキペディアより

データの大きさ

テラ：10^{12}，ペタ：10^{15}　エクサ：10^{18}　ゼッタ：10^{21}

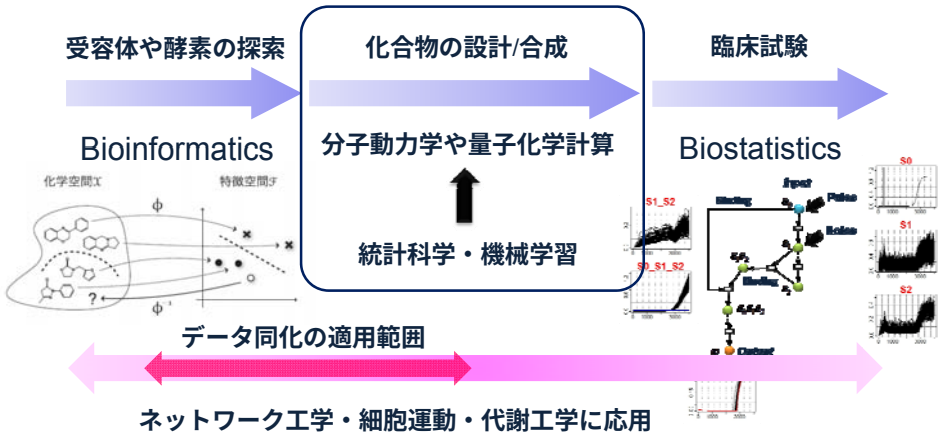
1TB （8Tbit）のハードディスク：　12, 800円
100 TBit：　米国議会図書館の情報の総量（全てがデジタル化された場合）
1エクサビット：　世界の印刷物の情報の総量
1ゼタビット：　Googleが推計した2009年6月の全世界のインターネットにおける情報の総量

インターネット上のデータ量は表層部分。大氷山の一角以下。
[ビジネス]
GREEのログ：　TB級/日

[科学]
ALMA望遠鏡：　数百TB/年　（2TB/日）

ビッグデータと創薬のかかわり

研究開発費: **製薬企業大手1社当り**1,274**億円** (**1成分**)
開発期間: 9**年～**17**年**



富を産む仕組み

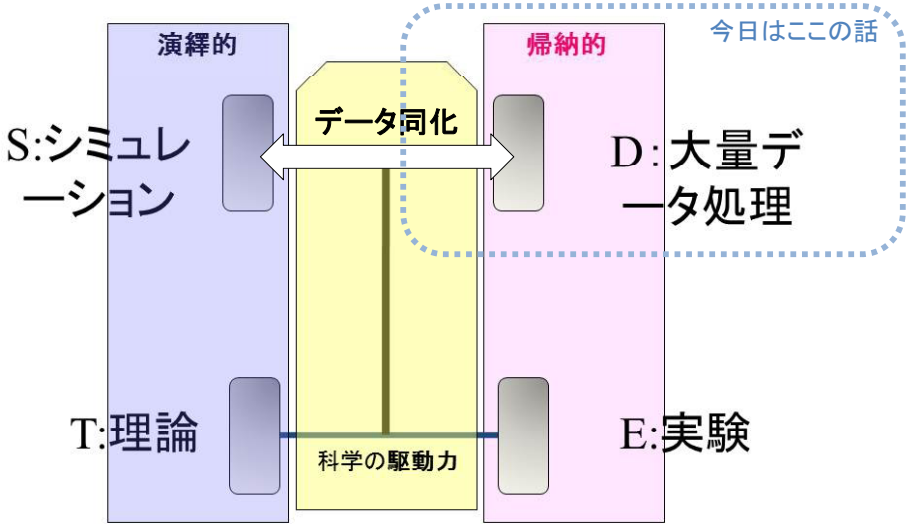
前世紀： 物質(「もの」)を均質に大量に生産するシステム

21世紀： 個人化された情報サービスを提供するシステム

個人をターゲットにした商品・サービスの提供を効率的に行えるシステム

“コ”ー個人，個性，個別，固有ーが大切！

つなぐ：データ同化



Big data techniques

TECHNIQUES FOR ANALYZING BIG DATA

A/B testing
Association rule learning
Classification
Cluster analysis
Crowdsourcing
Data fusion and data integration
Data mining
Ensemble learning
Genetic algorithms
Machine learning
Natural language processing
Neural networks
Network analysis
Optimization
Pattern recognition
Predictive modeling
Regression
Sentiment analysis
Signal processing
Spatial analysis
Statistics
Supervised learning
Simulation
Time series analysis
Unsupervised learning
Visualization

統計
機械学習
データマイニング
最適化
計算科学その他

ビッグデータと新NP問題

■ 1パラメータの値を、0～9の値から定める。

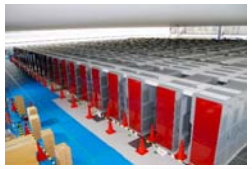
離散最適化問題

$$\max . f(\theta) \quad \theta' = (\theta_1, \ldots, \theta_p)$$

パラメータ数が2個（p=2）なら、10 x 10 =100 通り計算すればよい。

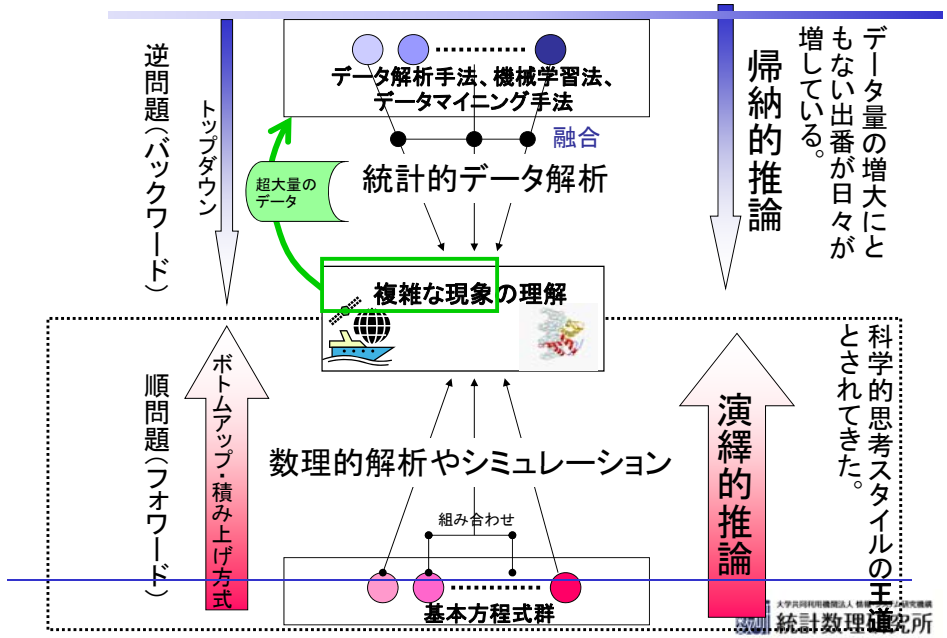
p=10　10^{10}：100億　（世界の人口が約70億人）
p=15　10^{15}：1000兆　（「京」の計算速度は8000兆回/秒）
p=20　10^{20}：1垓（がい）　（ルービックキューブの全パターン数の約2倍）

10^{150} 将棋のゲーム木の大きさ
10^{365} 囲碁のゲーム木の大きさ
Wikipediaより



スパースなデータ空間を N（サンプル数）の増大だけでカバーする（埋める）のは原理的に無理。**データ空間の中で構造を見つける方法が鍵。**

帰納的アプローチ



ベイズの定理がなぜ今役立つのか？４つの理由

イギリスの牧師・数学者（1702 – 1761年）
1763年に発見

x：興味のある対象

y：データ

2. 対象の特徴をとらえるセンサー性能の向上
高精度センサーのコモディティ（日用品）化

4. 高速（無線）インターネット網の整備

ベイズの反転公式

$$p(x \mid y) = \frac{p(y \mid x)\,p(x)}{\sum p(y \mid x)\,p(x)}$$

1. 膨大な数の積分（和）操作には高速な計算機が必要
コンピュータの性能向上

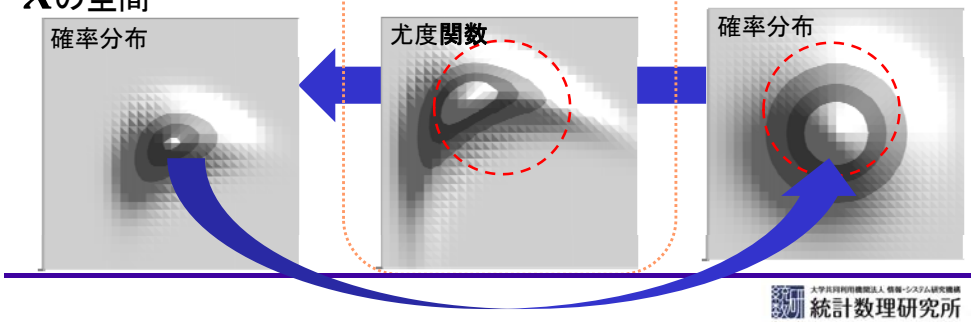
3. 対象の細かい情報を不確実性を含めて数値化。個人の情報を網羅的に収集
ストレージの廉価化

ベイズの定理と情報循環

$$p(x \mid y) = \frac{p(y \mid x) \cdot p(x)}{p(y)} \propto p(y \mid x) \cdot p(x)$$

Posterior	Likelihood	Prior
Improved knowledge about values of x	Feasibility of realization of y for given x	Belief about values of x

xの空間



ビッグデータをとりまく問題

・人材育成（人材争奪戦）

・法体系整備、プライバシー

統計数理研究所

統計学科を保持しない唯一の国が日本

- データを分析（解析）し、意志決定を行うためのプロフェッショナルを系統的に育成する機関がーつしかない。
- 第4の科学を研究する教育機関（組織）が統数研以外稀少
- データ分析結果に裏打ちされた優れたビジネスモデルこそが国際競争力を産む
- 演繹至上主義（「“真理”の探究」偏重）ー本教育の弊害。

第2次世界大戦以降統計学科が配置：
OECD諸 国、中国、韓国、台湾、香港、インド、バングラディシュ、シンガポール、南アフリカなどの主要大学

米国： 統計学 科自体が分野別に細分されており、生物統計学科、医学統計学科といった学科が存在

韓国： 統計関連学科としては、統計学科が16、情報統計学科が19、応用統計ないしは応用統計情報学科が5、生物統計学科1、保険数理統計学科が1である。これに加えて、統計関連学科として位置づけられている、Data Business学科、e-business学 科なども存在

中国： 2000年以降積極的な統計家育成が興り、2005年 現在で統計学科の数は161、学生総数2,5000人 であり、この他にも統計専門学校が300校設置