ビッグデータの活用における落とし穴

樋口知之　（情報・システム研究機構　統計数理研究所）

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

ゲノム解析を題材にとっても

理論・方法

- 情報量規準
- 時系列解析法
- 社会調査法
- 数量化理論
- 多変量解析

データ環境の変化とともに変容

- ベイズ統計
- MCMC, 粒子フィルタ
- アンサンブル学習器
- スパースモデリング
- カーネル法
- メタアナリシス
- データ同化

鎌谷先生@理研のスライドを改変

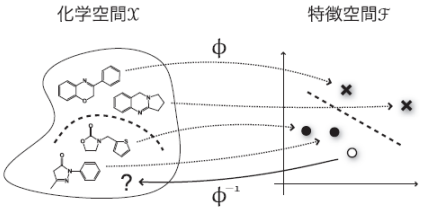
ビッグデータと創薬のかかわり

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

受容体や酵素の探索 → 化合物の設計/合成 → 臨床試験

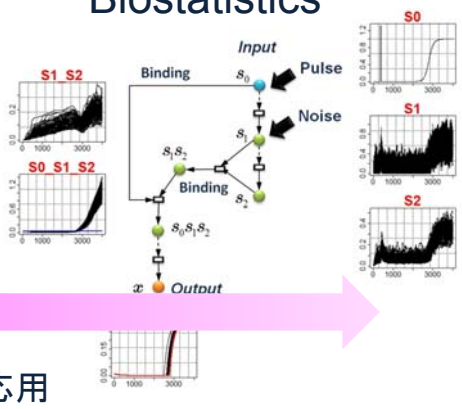
Bioinformatics



分子動力学や量子化学計算

統計科学・機械学習

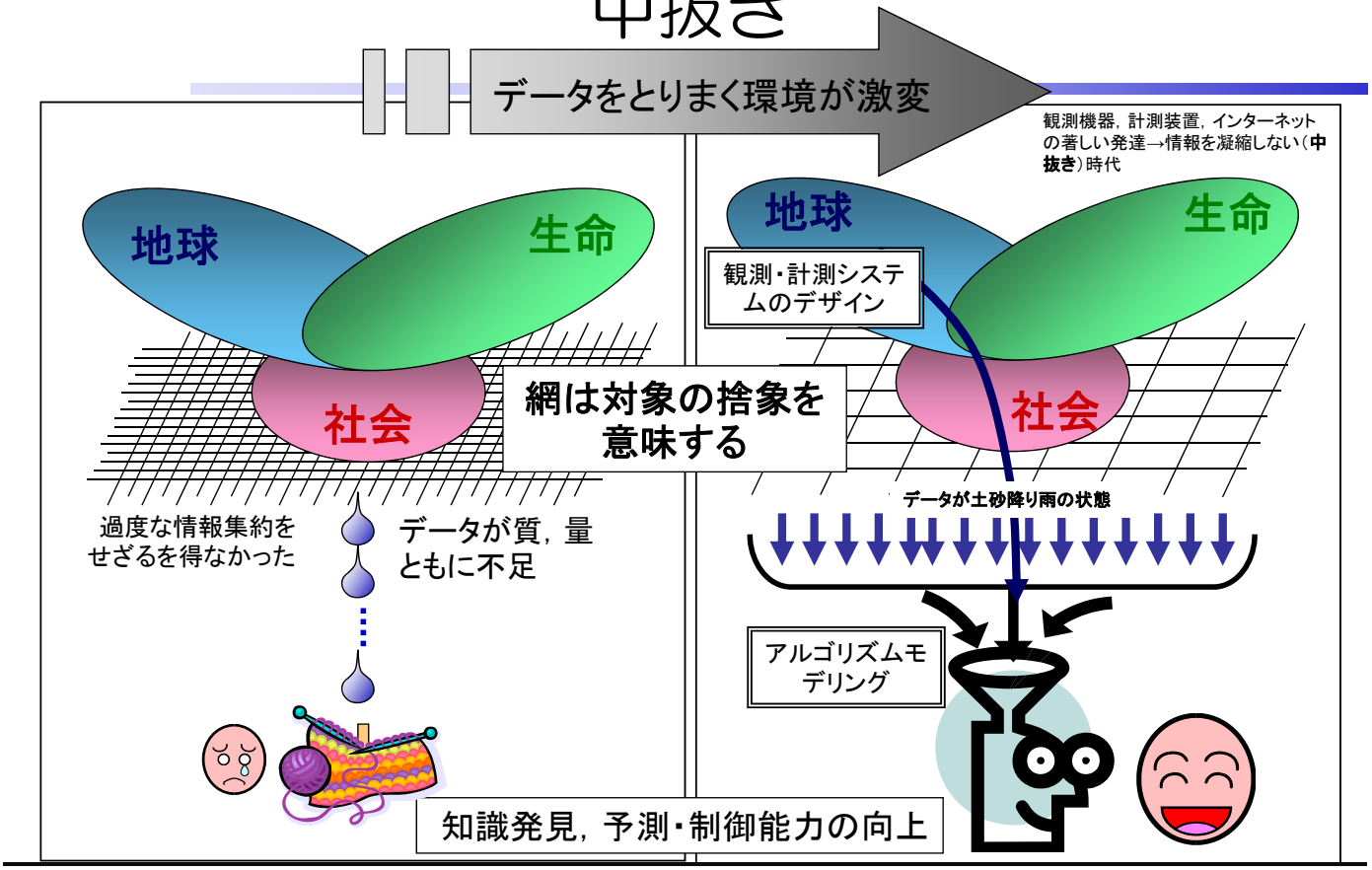
Biostatistics



ネットワーク工学・細胞運動・代謝工学に応用

吉田@統数研のスライドを改変

中抜き



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."

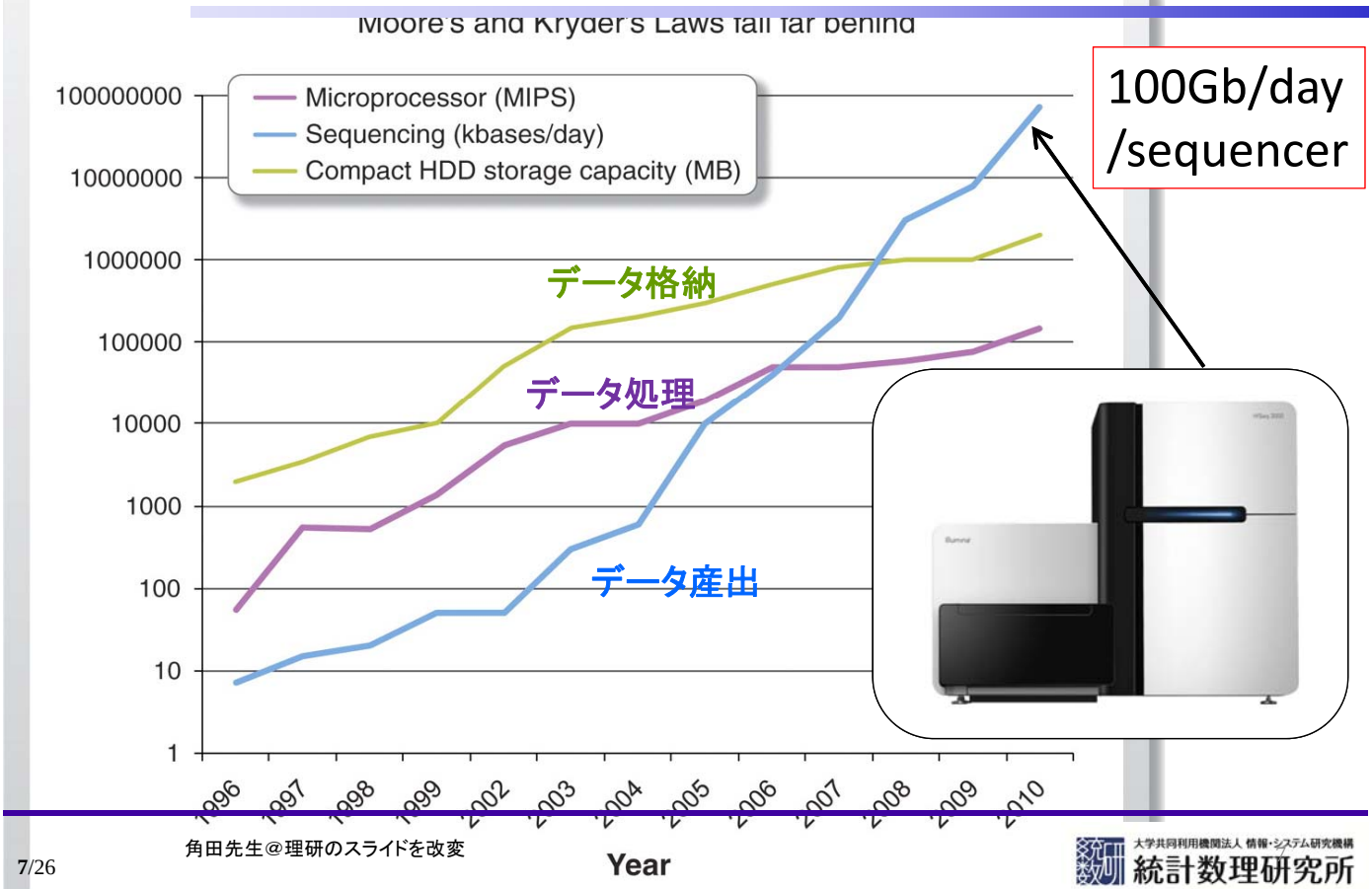
http://www.nsf.gov/news/news_images.jsp?cntn_id=123607

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

データ中心主義の時代へ



角田先生＠理研のスライドを改変

富を産む仕組みも変わった！

前世紀：　物質（「もの」）を均質に大量に生産するシステム

↓

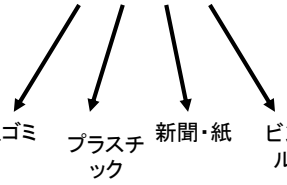
21世紀：　個人化された情報サービスを提供するシステム

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！

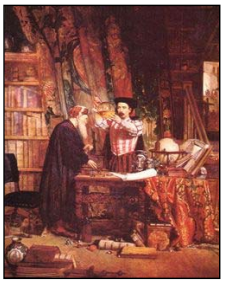
ビッグデータは巨大なゴミ箱？

ビッグデータの実際は、**そのまま**だと単なる屑の山

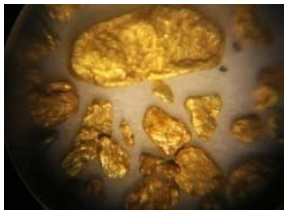


分別、整理することで

1．マイニングは錬金術師でしょ？
データ解析への懐疑的態度

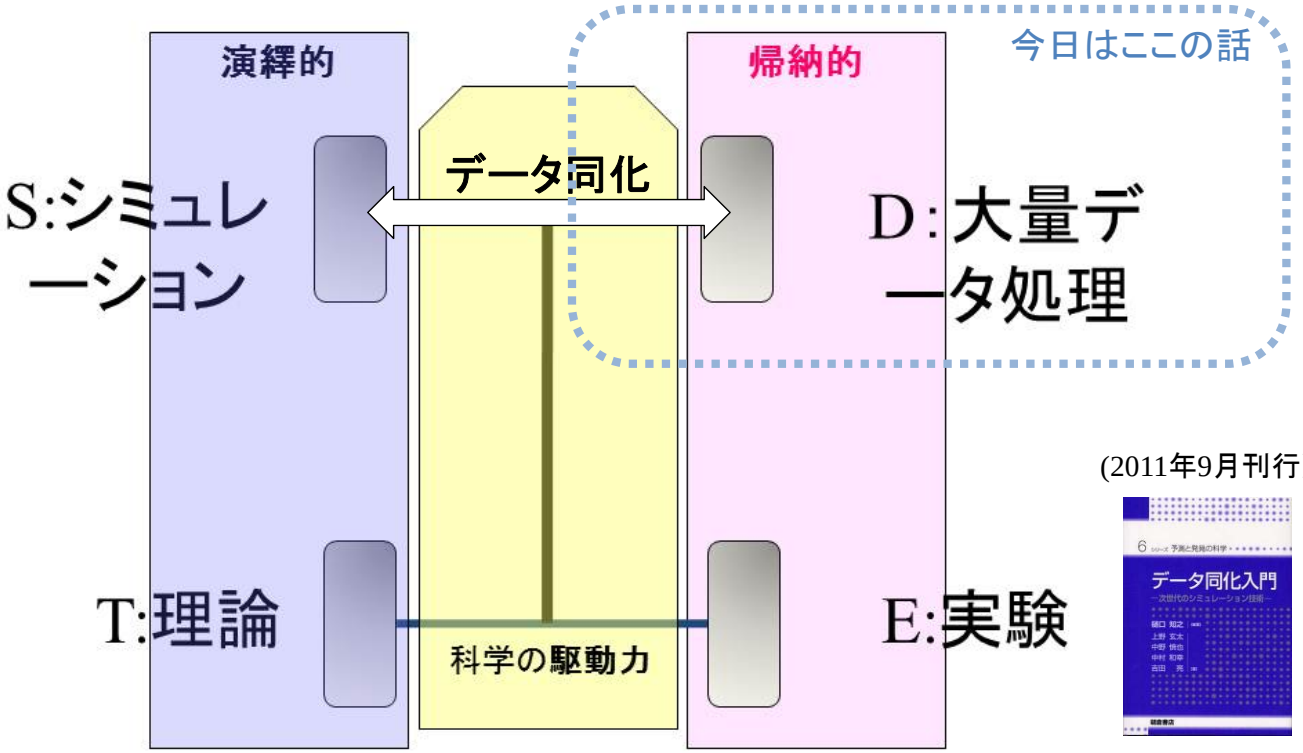


2．砂金探しをいつまで続ける？
エキスパートへの過度な依存



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

つなぐ：データ同化



(2011年9月刊行)

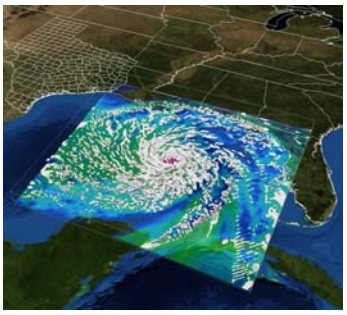
大学共同利用機関法人 情報・システム研究機構 統計数理研究所

NSF Leads Federal Efforts In Big Data

March 29, 2012

On March 29, the Federal Government held a webcast Federal government science leds from OSTP, NSF, NIH, DOE, DOD, DARPA and USGS outlined how their agencies are engaged in Big Data research, followed by a panel of thought leaders from *academia and industry*, moderated by Steve Lohr of the New York Times.

About Big Data: **Researchers in a growing number of fields are generating extremely large and complicated data sets, commonly referred to as "big data."** A wealth of information may be found within these sets, with enormous potential to shed light on some of the toughest and most pressing challenges facing the nation. To capitalize on this unprecedented opportunity--to extract insights, discover new patterns and make new connections across disciplines--we need better **tools to access, store, search, visualize and analyze** these data.



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

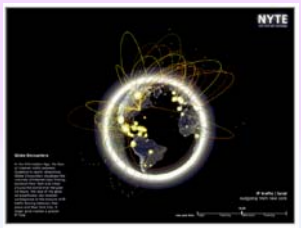
Big data techniques and technologies

- **TECHNIQUES FOR ANALYZING BIG DATA**

統計科学、数理工学

- **VISUALIZATION**

Tag Cloud



Spatial Information Flow

McKinsey Global Institute
June 2011
Copyright © McKinsey & Company
www.mckinsey.com/mgi

- **BIG DATA TECHNOLOGIES**

Big table, Business intelligence, Cassandra, Cloud computing, Hadoop, MapReduce, Relational database, Stream processing

計算機科学、情報工学

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

ビッグデータ環境下における研究開発推進の鍵

・個人化技術 (*Personalization*)

落とし穴1： 新しい「*NP*問題」（次元の呪い）

落とし穴2： 相関と因果

（超高次元の情報空間内の構造探索とモデル化）

・帰納的推論と機能のモデル化

落とし穴3： 物理帝国主義観からの脱却

（ニュートンパラダイムからのシフト）

落とし穴1：ビッグデータと新*NP*問題

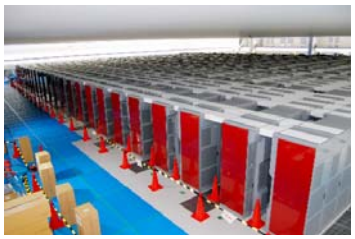
■ 1パラメータの値を、0～9の値から定める。

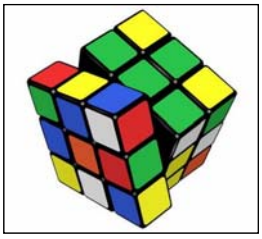
離散最適化問題

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

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）

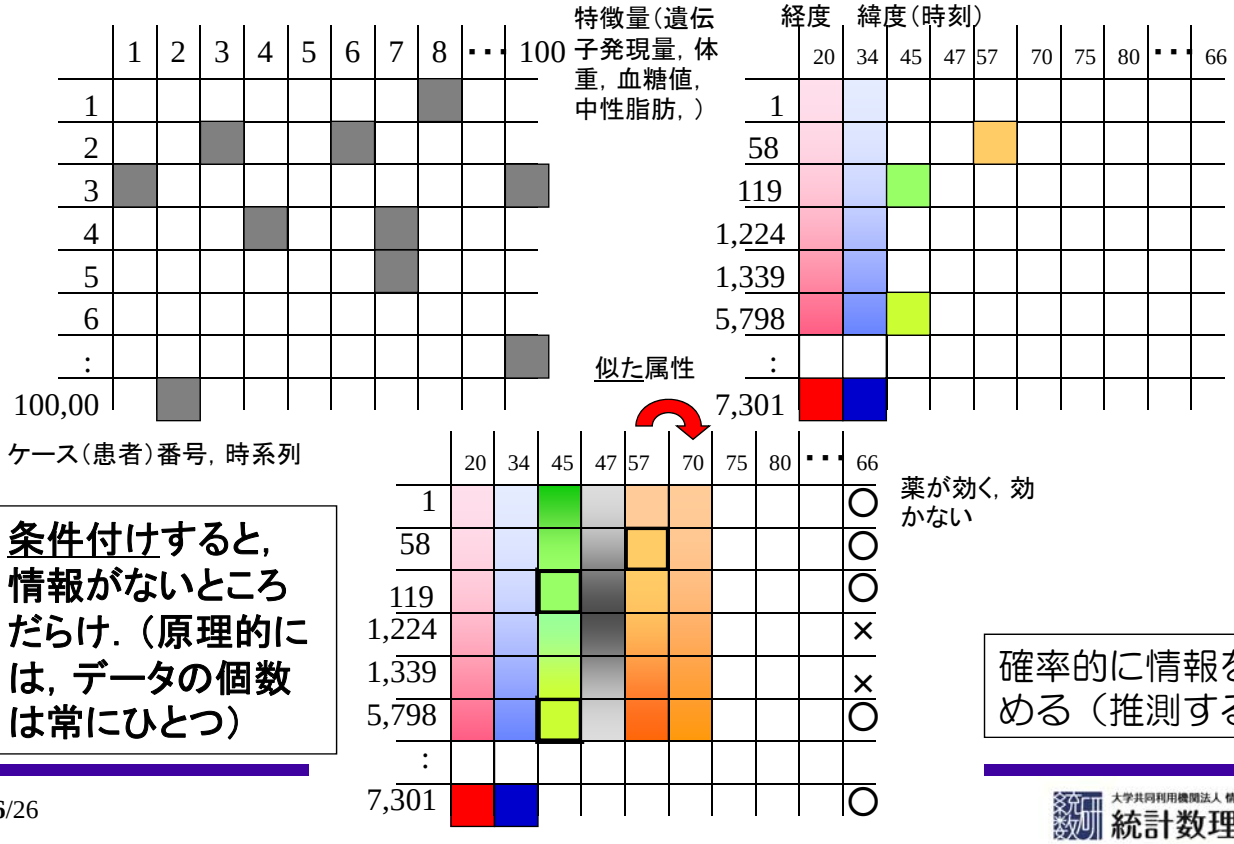




10^{150} 将棋のゲーム木の大きさ
10^{365} 囲碁のゲーム木の大きさ　Wikipediaより

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。**

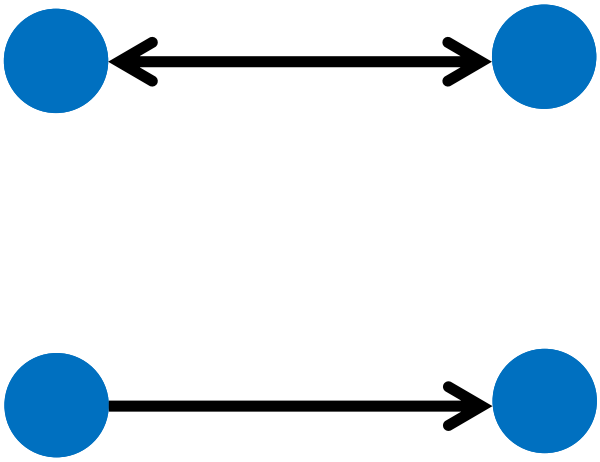
データの有限性→情報の欠損



条件付けすると，情報がないところだらけ．（原理的には，データの個数は常にひとつ）

確率的に情報を埋める（推測する）

落とし穴2：列挙処理、相関と因果



見えないものをビッグデータで推量する

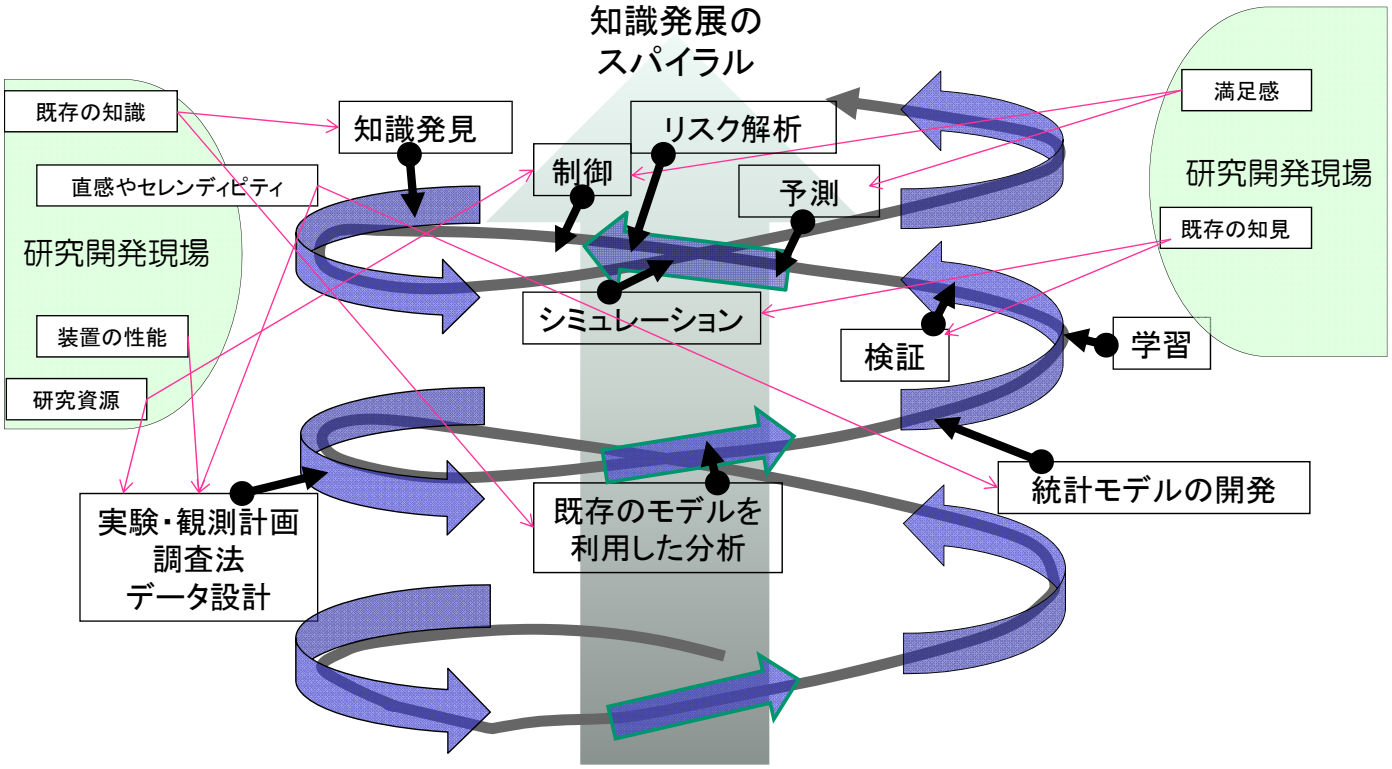
わからないもの，見えないもの

- 人間・生活活動に関連する大量大規模データの出現
- 不十分ではあるが，一人一人の行動にいたる考え方を間接的に捉えることが可能に.

構造が確認されている確率的な機構という特殊な場合を除き、期待の構成の仕方は**我々の持つ知識や経験の使い方に大きく依存**する。したがって、唯一無二の真の構造のようなものは存在しない。…..したがって、我々はより良いモデルの探求を通じて、常に未知の状態にある究極的な真理あるいは真の構造に迫るのである。

赤池弘次「時系列解析の心構え」、朝倉書店(1995)

知識循環と永続的なモデルの改良



ビッグデータ環境下における研究開発推進の鍵

・ 個人化技術 (*Personalization*)

落とし穴1： 新しい「*NP*問題」（次元の呪い）

落とし穴2： 相関と因果

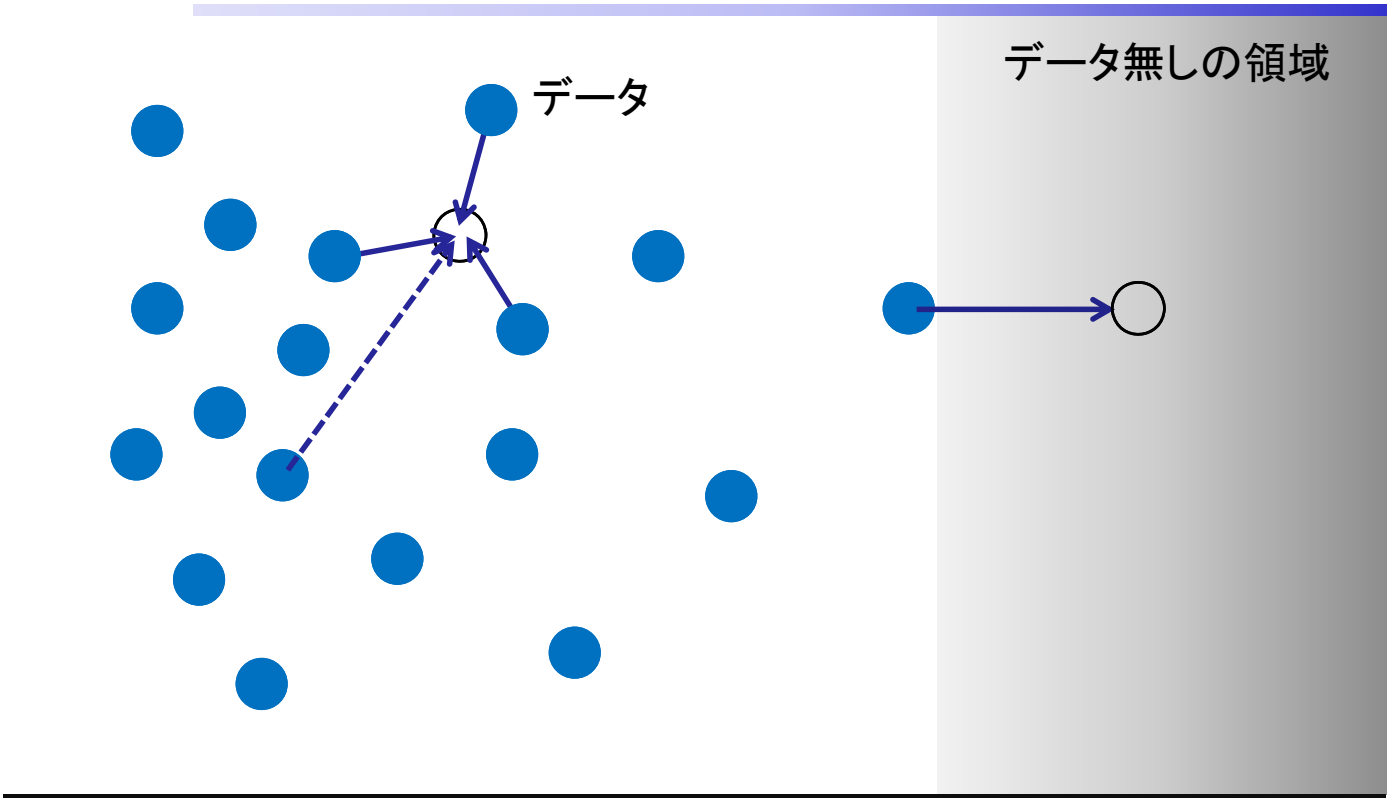
（超高次元の情報空間内の構造探索とモデル化）

・ 帰納的推論と機能のモデル化

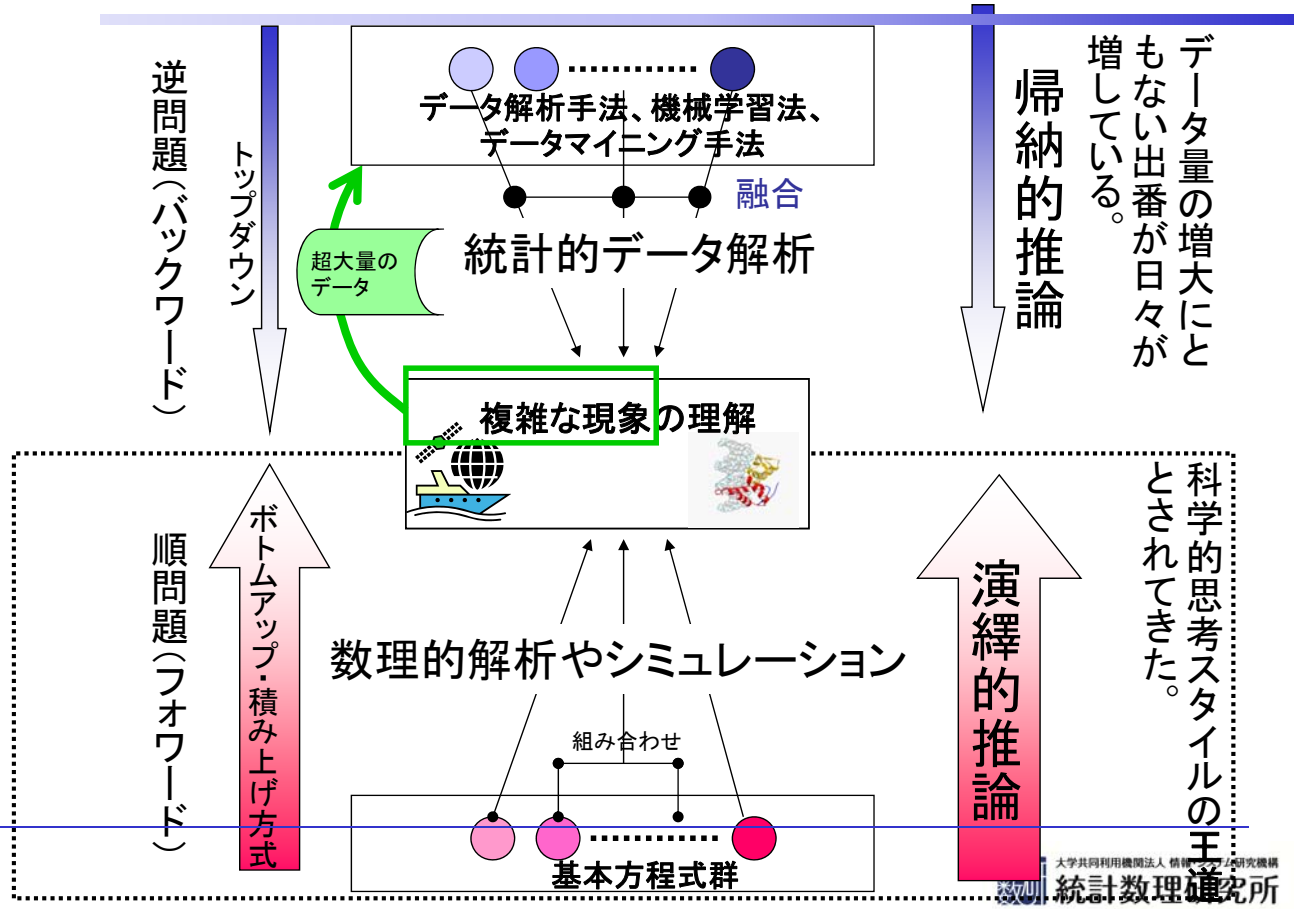
落とし穴3： 物理帝国主義観からの脱却

（ニュートンパラダイムからのシフト）

落とし穴3：内挿と外挿問題



帰納と演繹の両方が大切



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

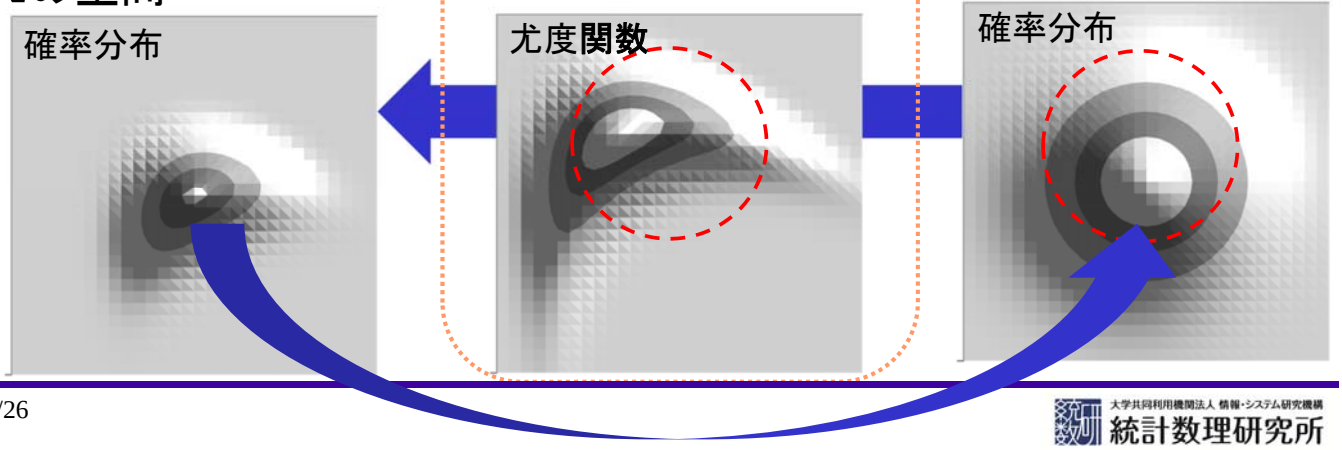
3. 対象の細かい情報を不確実性を含めて数値化。個人の情報を網羅的に収集
ストレージの廉価化

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

データリテラシーは大丈夫？

統計学科を保持しない唯一の国が日本

たった一つしか

- 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一つしかない。
- 第４の科学を研究する教育機関（組織）が統数研以外皆無
- 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」）一本教育の弊害。

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