ビッグデータ時代における
科学的課題への新たなアプローチ法

e-サイエンス：超大規模実問題に挑戦するアルゴリズムと計算技術

西尾 章治郎
大阪大学大学院情報科学研究科

2012年9月4日

計算能力の高速化およびネットワークの広帯域化

計算能力の高速化： 「京」は、1秒間に1京回（10の16乗回）の演算性能を実現

従来の狭帯域ネットワーク

小さいメッセージを複数回交換

ネットワーク技術の発展に伴い、ネットワークの帯域幅が拡大すると…?

1Gbps

10Gbps

100Gbps

広帯域ネットワーク

ある程度大きな情報をまとめて交換

1G:10の9乗（10億）

10000人×10KBの人事データ（100MB）を実効速度100Mbpsのネットワークで転送すると約1秒程度の時間で送れる

記憶装置の大容量化

ワークステーションの記憶装置の大容量化

1988年：1GBで**約100万円**
（ギガ＝10の9乗、つまり、10億。本を1000冊分記憶）

USBメモリ，光学ディスク，HDD

2012年：USBメモリ 64GB **約8000円**
BD-R 25GB **50円～/枚**
DVD-R 4.7GB **15円～/枚**
HDD **5円～/GB**

情報通信技術の驚異的な発展

情報通信技術の
急速な発展…

しかし、
どうのように
活かすの?

ついて
行くのが
大変だ～！

性能指標
100億倍

価格
10万分の1程度

情報通信技術

1960年からの半世紀の間に、
価格性能比で
1兆倍以上の変化!

**このような急激な変化を起こした産業は、
他に類を見ない**

交通機関

**ここ50年間で速度や燃費が
数倍に向上したに過ぎない**

それってホントに幸せ？

何かおかしい…

安浦寛人先生（九州大学）
の試算による

e-サイエンス: 第4の科学パラダイム

数千年前： 経験科学

- 自然現象を解明する科学

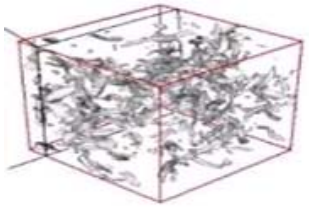
数百年前： 理論科学

- ニュートンの法則、マクスウェルの等式など、理論中心の科学

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

数十年前： 計算科学（シミュレーション科学）

- 複雑な現象をシミュレーションにより予測・再現する科学



現在： e-サイエンス または データセントリックサイエンス

- 理論、実験、および シミュレーションを統合する
- データ検索、データマイニングを用いる
  - 実験器具からのデータを取得
  - シミュレーションからデータを生成
  - ソフトウェアによって処理
  - 科学者は、データベースやファイルを分析

科学の方法論のパラダイムシフト
(Jim Gray博士のリーダーシップ)

マイクロソフト社 Tony HAY氏講演より

e-サイエンスの具体的形態 （あるべき姿：実際はこれらの複合体）

1. ネットワーク上での「計算科学」連携システム

ex. ペタコン／情報基盤センター・スパコン／キャンパスレベル・クラスタマシン等
これからは、アカデミッククラウド環境の構築が課題

2. 実験リソース（特殊装置）の高速ネットワーク上での共用

ex. 電子顕微鏡、化学実験装置、SPring-8

3. 観測データリアルタイム連携（センサネットワーク型）

ex. 天文（バーチャルオブサバトリ）、地震観測ネット

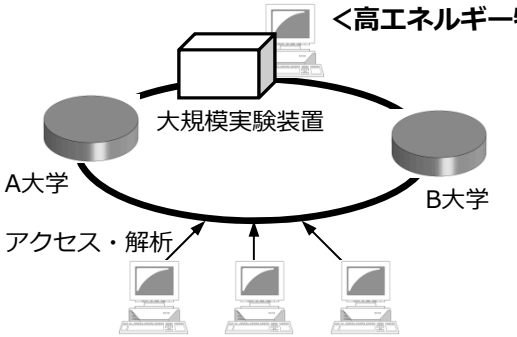
4. サイエンスデータベース、学術コンテンツ等の（データ）共用

ex. 統合データベース、電子ジャーナル、次世代学術コンテンツ基盤

e-サイエンスの例： スーパーSINETを利用した代表的な研究成果

広域分散データ解析システム

<高エネルギー物理学等>

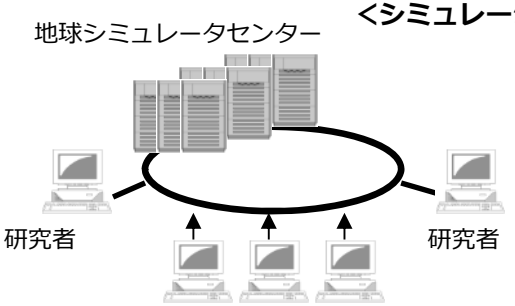


先端研究実験データのリアルタイム同時並行解析の実現

高エネルギー加速器研究機構で行われるBelle実験の大規模データを、東北大、東大、東工大、名大、阪大で解析し、「ＣＰ保存則の破れ」を検証。

連結階層シミュレーション マルチスケールシミュレーション

<シミュレーション科学等>

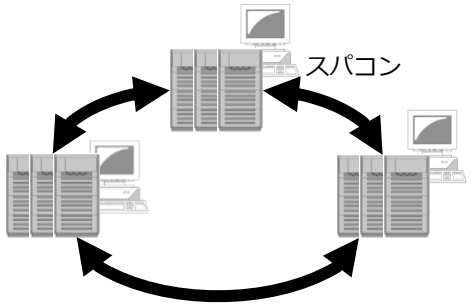


地球シミュレータセンターを高速ネットワークで利用し、新しいシミュレーション手法の研究開発を実施

地球シミュレータセンターを高速ネットワークで利用し、ヒートアイランドなどの都市現象の予測などの新しいシミュレーション手法を開発し、都市設計の提言などに利用
（海洋開発研究機構）

大規模設計シミュレーション

<材料・ナノテクノロジー等>

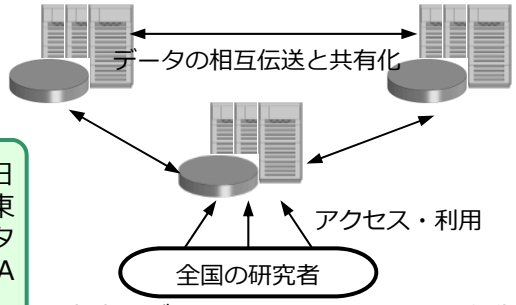


複数のスパコンをネットワークで結合し、材料設計等のための大規模シミュレーションを実施

東北大、東大、九大、分子科学研究所のスーパーコンピュータを直接連結し、超大規模シミュレーション計算を実行。水素吸収水化合物の構造の最適化や水素分子と水分子の結合状況の確定に成功。

広域分散データベース整備と遠隔利用

<バイオインフォマティクス等>

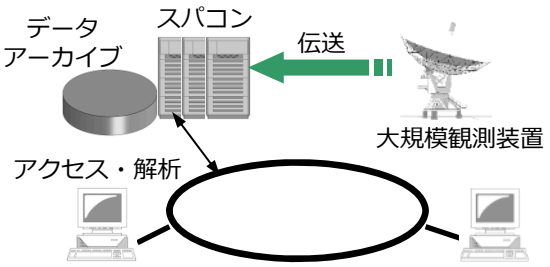


研究成果データをネットワークで共有化し、全国の研究者に公開

国立遺伝学研究所で運営する日本DNAデータバンクの情報を東大、京大で共有し、大量データを更新。米国、欧州と国際DNAデータバンクの共同構築。

観測データの遠隔解析

<天文学・宇宙研究>



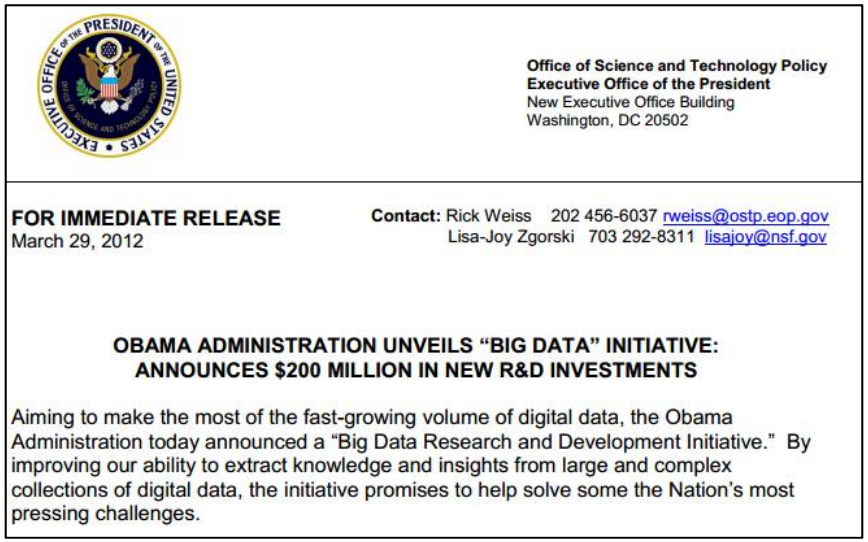
大規模な観測データを遠隔地から解析し研究に活用

遠く離れた電波望遠鏡を超高速回線で結合し、これまでにない高い観測感度を達成する世界最長の光結合型ＶＬＢＩ（超長基線電波干渉計）の実現
（国立天文台）

米政府 “ビッグデータ”の利活用を目的とした研究開発イニシアチブを発表

オバマ政権は新規に＄200MをBig Data R&D Initiativeに投じることを発表

- 大規模で複雑なデジタル・データから知見を引き出す能力を高めることによって、国家の喫緊の課題を解決するために役立てることが目標
- 今後、科学的発見、環境・生物医学研究、教育、国家安全保障のためにBig Dataを活用する能力が変革することを狙う



Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a “Big Data Research and Development Initiative.” By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation’s most pressing challenges.

米ホワイトハウス科学技術政策局（OSTP: Office of Science and Technology Policy）
2012年3月29日発表

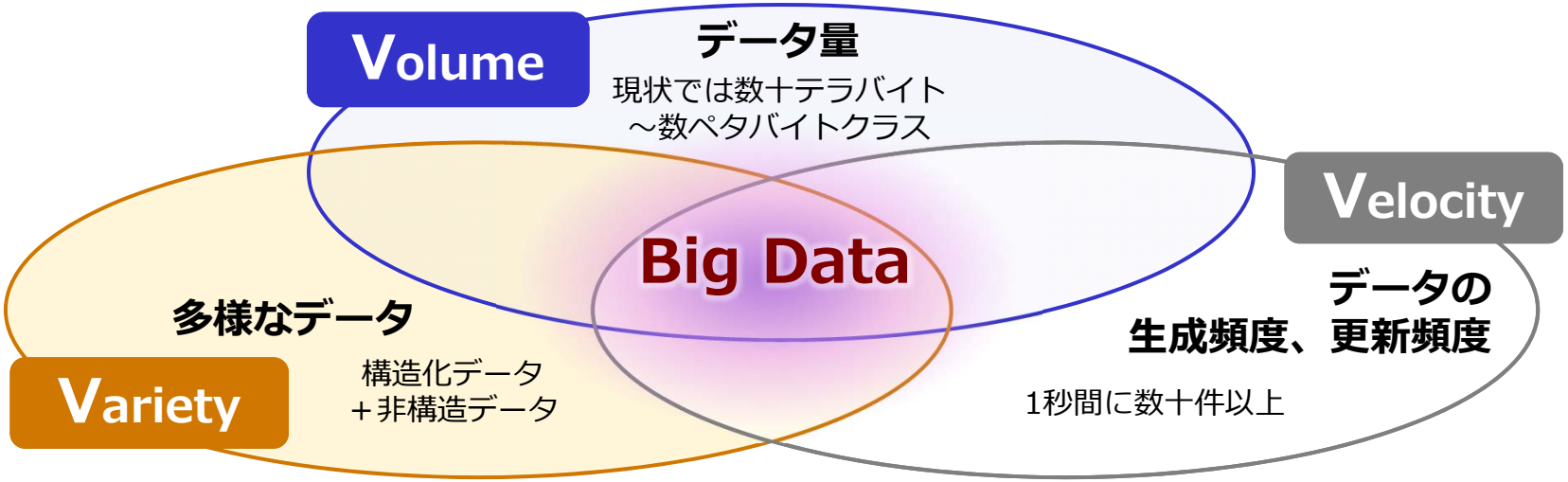
ビックデータとは？

狭義の定義

ビックデータとは、既存の一般的な技術では管理することが困難な大量のデータである。

広義の定義

ビックデータとは、3Vの面で管理が困難なデータ、および、それらを蓄積・処理・分析するための技術、さらに、それらのデータを分析し、有用な意味や洞察を引き出せる人材や組織を含む包括的な概念である。

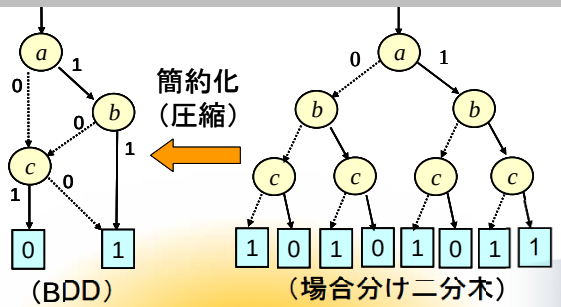


城田真琴著：ビックデータの衝撃（東洋経済新報社）

計算パワーだけに頼るのでなく、アルゴリズム基盤の革新が要請されている

「京」「TSUBAME」をはじめとするスーパーコンピュータを最大限に活かし、高速ネットワーク上の膨大な実データを分析・解析するe-サイエンスの基盤となるアルゴリズム科学にブレークスルーが求められている

最先端理論 （Algorithm Theory）



有機的な結合

超大規模実データ
(Big Data)

最新計算技術
（Computation）

革新的アルゴリズム（最適化、列挙、探索等）による社会システムのデザイン

- データマイニング、人工知能、機械学習、統計、HPCなどの分野と結合し、**新しい学問体系の創生と人材育成**
- 物理、化学、生物などの科学、土木、建築、機械などの工学、環境・エネルギー、交通、経済・経営の**諸分野への適用**
- 地球的規模の諸問題(環境、エネルギー、バイオロジー)および**社会的危機管理**(防災、避難および復興計画の策定）への対応

日本で推進することの優位性と緊急性

機が熟す　**日本的な研究スタイルの優位性を発揮**

ERATO 湊離散処理構造処理系プロジェクトなどを契機に、
加藤直樹先生のリーダーシップによるコミュニティの纏まり

コミュニティの総力を結集できる状態

戦略性　**e-サイエンス、ビックデータ時代の到来**

緊急の課題（災害対策等）や新分野での革新的アルゴリズム
開発と技術覇権の獲得を目指す

世界をリードする立場を確保する

緊急性　**東日本大震災の発生**

大規模災害に対する社会的危機管理や耐災害社会設計が
極めて重要

1日の遅れが多大な社会的損失につながる

日本学術会議マスタープランの策定(1)

提言

学術の大型施設計画・大規模研究計画
－企画・推進策の在り方とマスタープラン策定について－

2010年3月17日 日本学術会議 科学者委員会 学術の大型研究計画検討分科会

- 7分野43の研究計画をリストアップ（各計画を科学的視点で評価）
  - **A計画: 大型施設計画**
    大型の研究施設・設備を建設・運用する計画（建設費１００億円超）
  - **B計画: 大規模研究計画**
    大規模な研究基盤・ネットワークの構築やデータ集積等を行う計画
    （初期投資と運営費を併せて数十億円以上）
- 情報関連計画は、「情報インフラストラクチャ」分野にA計画１件、B計画1件採択

文部科学省の科学技術・学術審議会に大型プロジェクト作業部会

- マスタープランに盛り込まれた43計画のヒアリング（4日間）
- ロードマップの内容検討

＜評価の観点＞
- 研究者コミュニティの合意、計画の実施主体、共同利用体制、計画の妥当性
- 緊急性、戦略性、社会や国民の理解

- 〔パブリックコメント実施（2010年9月10日～10月12日）〕
- 2010年10月27日　ロードマップを含む「審議のまとめ」策定・公表

日本学術会議マスタープランの策定(2)

文部科学省 最先端研究基盤事業

- 2010年度開始事業（予算額：300億円程度）
- 国際的な頭脳循環の実現に向け、国内外の若手研究者を惹きつける研究基盤の整備を強化・加速するため、研究ポテンシャルが高い研究拠点において、最先端の研究成果の創出が期待できる設備を整備するとともに、運用に必要な支援を行う事業
- ロードマップ掲載計画のうち、9計画が採択

マスタープランの軽微な見直し

報告　学術の大型施設計画・大規模研究計画 マスタープラン2011

2011年9月28日 日本学術会議 科学者委員会 学術の大型研究計画検討分科会

- 7分野46の研究計画をリストアップ
- 情報関連計画は、「情報学」分野のB計画に3件採択
  その内の一つ： e-サイエンスに向けた革新的アルゴリズム基盤

文部科学省の科学技術・学術審議会に大型プロジェクト作業部会

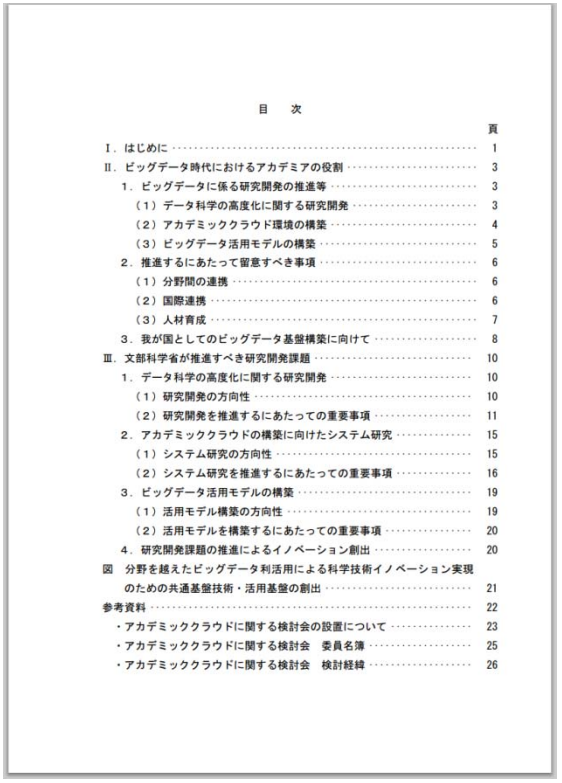
- マスタープランに盛り込まれた46計画の内、新たな計画などを15計画をヒアリング（4日間）
- ロードマップの内容検討

＜評価の観点＞ 前回同様７項目の視点

- 〔パブリックコメント実施（2012年4月11日（水）～５月７日（月））〕
- 2012年5月28日　ロードマップの改訂の決定・公表

文部科学省のアルゴリズム関係の重要性の認識

ビッグデータ時代におけるアカデミアの挑戦
～アカデミッククラウドに関する検討会 提言～



目　次

2012年7月4日文部科学省 アカデミッククラウドに関する検討会

Ⅲ．文部科学省が推進すべき研究開発課題

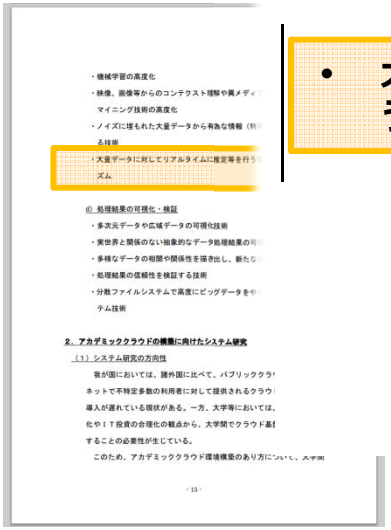
1．データ科学の高度化に関する研究開発

(2) 研究開発を推進するにあたっての重要事項

② ビッグデータ利活用のための共通基盤技術の開発

c) データの分析・処理

（15頁）



・機械学習の高度化
・映像、画像等からのコンテクスト理解や異メディ
マイニング技術の高度化
・ノイズに埋もれた大量データから有為な情報（特
る技術
・大量データに対してリアルタイムに推定等を行う
ズム

d) 処理結果の可視化・検証
・多次元データや広域データの可視化技術
・実世界と関係のない抽象的なデータ処理結果の可
・多様なデータの相関や関係性を描き出し、新たな
・処理結果の信頼性を検証する技術
・分散ファイルシステムで高度にビッグデータを中
テム技術

2．アカデミッククラウドの構築に向けたシステム研究
（1）システム研究の方向性
我が国においては、諸外国に比べて、パブリッククラ
ネットで不特定多数の利用者に対して提供されるクラウ
導入が遅れている現状がある。一方、大学等においては、
化やＩＴ投資の合理化の観点から、大学間でクラウド基
することの必要性が生じている。
このため、アカデミッククラウド環境構築のあり方に

- 15 -

（特異点）と自動抽出する技術

- **大量データに対してリアルタイムに推定等を行う等の革新的アルゴリズム**