SDS-1

ビッグデータ時代の方法論：統計科学とデータサイエンス

北川 源四郎

November 2017

Statistics & Data Science Series back numbers:
http://www.mims.meiji.ac.jp/publications/datascience.html

ビックデータ時代の方法論：統計科学とデータサイエンス

Methodologies in the Big Data Era: Statistical science and data science

北川 源四郎
明治大学
先端数理科学インスティテュート

2017年5月27日 明大中野キャンパス
特別講演会「統計科学のフロンティア」

プロフィール： 北川源四郎

現職： 明治大学 先端数理科学インスティテュート 所員
統計数理研究所 および総合研究大学院大学 名誉教授
日本学術会議会員（第3部）

略歴： 東京大学理学系研究科数学専攻博士課程中退（1974），理学博士。
統計数理研究所研究員，助教授，教授，所長（02－11），情報・システム研究機構長（11－17）を歴任。この間、タルサ大学客員助教授（80-81)、合衆国商務省センサス局研究員（81-82）、総合研究大学院大学助教授・教授（88-11)、東京大学経済学研究科助教授（88-91)、日本銀行金融研究所客員研究員(96-98)。

研究分野： 時系列解析（非定常モデリング，粒子フィルタ）
統計的モデリング（情報量規準GIC, EIC，ベイズモデリング）
船舶のオートパイロット、経済時系列の季節調整、地震波の自動処理、信号抽出、強風予測（JR列車安全運行システム）などの応用研究

所属学会： ASA (American Statistical Association: Fellow), ISI (International Statistical Institute), IASC (International Association for Statistical Computing), 日本統計学会,日本数学会, 応用統計学会, 人工知能学会, 計算機統計学会，日本船舶海洋工学会, 日本地震学会, 計測自動制御学会, 応用経済時系列研究会, 日本金融・証券計量・工学学会

過去の学会活動： 日本統計学会会長，統計関連学会連合理事長, ISI Councilor, IASC Councilor

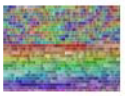
主な著書： 情報量統計学(共立出版1983),時系列解析の実際I,II（朝倉書店1994)、時系列解析の方法（朝倉書店1998),情報量規準（朝倉出版2004）、時系列解析入門（岩波書店、2005)
Akaike Information Criterion Statistics（D.Reidel 1986)
Smoothness Prior Analysis of Time Series (Springer, 1996)
Practice of Time Series Analysis (Springer 1998)
Information Criteria and Statistical Modeling (Springer 2008)
Introduction to Time Series Modeling (Chapman & Hall 2010)

アカデミック・ビッグデータ Academic Big Data

情報通信・計測技術の飛躍的発展
Dramatic development of ICT and sensor

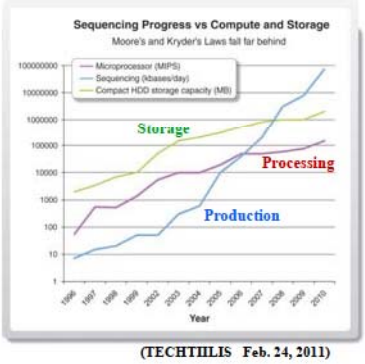
大量・大規模データの集積
Accumulation of big data

- Life Science: DNA, Micro-array data
- Marketing: POS data
- Finance: High frequency data
- Environmental Science
- Meteorology, Seismology
- Disaster Prevention
- Astronomy（Whole-sky CCD camera）
- High-energy physics（LHC）
- Linguistics :



ICT & Sensors

- Measurement, Sensors
- Internet
- Databases
- Super computers, GPU



(TECHTHLIS Feb. 24, 2011)

4

ソーシャル・ビッグデータ Social Big Data

人間の活動を精細・網羅的に記録し，デジタル化した結果
The result of accurately and comprehensively recording human

- Internet : Web, SNS, Mail, mobile phone
- Sensor data : 家電，自動車，GPS，RFID
- Transaction data : POS, Stock, Real estate
- Multimedia : Image, Sound,
- Log data: Software log, Life log


活用例

- Marketing（市場予測、顧客行動予測モデル）
- On-line shoppingにおける推奨機能
- Data-driven産業
- 医療・創薬・保健における個人化対応
- 社会インフラのスマート化
- Sensor data活用（防犯，防災，故障検出）
- Evidence Based Policy Making, Data Journalism

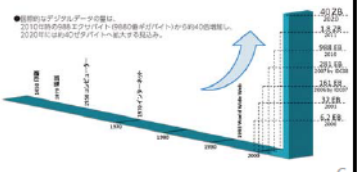


5

物質世界とは桁違いのデータ増加 Data growth is by far faster…

- **物質世界の増加** Increase in the real world
  - **人口:** Population　日本　4000万人（1900年）→ 12000万人（2000年）
    世界　10億人　（1800年）→ 60億人　（2000年）
  - **穀物生産:** Grain production　世界　6.31億トン(1950年) → 22.27億トン(2008年)
  - **工業生産:** Industrial production　日本　6.8兆円　（1955年）→ 290.7兆円　（2012年）
- **情報世界の増加** Increase in the cyber world
  - **ICT:**　Moore's law　2倍（1.5年），100倍（10年）
  - **Data**　Increase of data　6.2×10^{18}B（2000年）→ 4.4×10^{21}B（2013年）

	Growth rate	10 years GR	20 year GR
Population	1.01	1.09	1.20
Grain production	1.02	1.24	1.54
Industrial production	1.07	1.93	3.73
ICT	1.59	100	10000
Data	1.66	156	24000



6

ビッグデータのインパクト Impact of Big Data

- やがて，**超スマート社会**が出現する。
  Super smart society will appear before long.
  - **すべての産業がICT化**（金融，投資，資産運用，住宅設備，家電，自動運転，電気自動車，医療サービス，ヘルスケア，診断，手術，創薬，交通サービス）
  - **社会インフラのスマート化**（エネルギー，都市，交通システム，ビル管理）
  - **個別化（One to one）サービス**（医療，教育，マーケティング，情報提供）
- すべての研究は**データサイエンス化**する。
  All research will become data science.

7

ICTによる急速な変化 Rapid changes caused by ICT

	Past	Present
通信	手紙・電話・FAX	スマホ，メール，Twitter，LINE，Skype
検索	辞書・辞典，電話帳，地図，時刻表，ピア	翻訳サイト・Wikipedia，ネット地図，乗換案内，ぐるなび，
情報発信	新聞，雑誌，テレビ，ラジオ	Internet，You Tube，Twitter

一部Yahoo安宅氏のスライドから編集

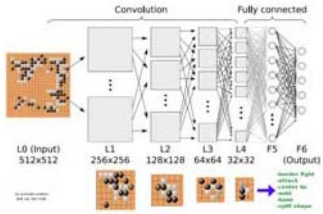
8

ICTによる急速な変化（２） Rapid changes caused by ICT

過去	現在
● **販売** Sales 個人商店，デパート，量販店，	コンビニ，ネット販売（Amazon, Yahoo）
	
● **物流** Logistics 調達⇨製造⇨流通⇨小売⇨顧客 （個別最適化）	SCM（サプライチェーン・マネジメント） （全体最適化）
	
● **犯罪捜査** Crime Investigation 聞き込み，刑事の勘，指紋照合	防犯カメラ，プロファイリング， DNA鑑定，指紋自動照合
	 

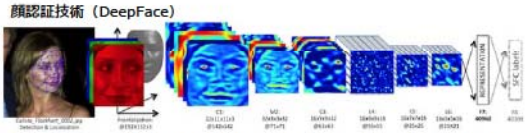
一部Yahoo安宅氏のスライドから編集

人工知能（特に深層学習の応用）が日々新しい成果を生み出す



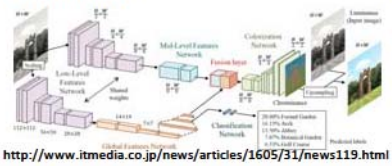
https://medium.com/udacity/deep-learning-nanodegree-foundation-program-syllabus-in-depth-2eb19d014533#.t3kngso0o

顔認証技術（DeepFace）



https://research.preferred.jp/2014/03/face-verification-deepface-and-pyramid-cnn/

自動彩色（早稲田大学）



http://www.itmedia.co.jp/news/articles/1605/31/news119.html

自動彩色

Source: http://blog.livedoor.jp/tsuchiyano/archives/52164365.html

ビッグデータが拓く世界 The world that the big data opens

日本学術会議提言「ビッグデータ時代に対応する人材の育成」
Science Council of Japan

1. データ駆動型産業の創出 Creation of data driven industries

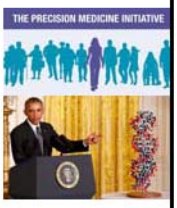
・地球規模で網羅的に収集したビッグデータを活用（グーグル）

2. 個人化サービスの実現 Realization of personalized service

・個別化医療・創薬（ゲノム情報や環境因子の統合活用による医療革新（QOL向上，財政課題緩和）

・One to Oneサービス提供（マーケティング，医療、教育，情報提供など）

・大量生産・大量消費による効率最大化から個人満足度最大化への転換



3. 1次産業・2次産業の効率化 Improve efficiency of primary industry and secondary industry

多数のセンサーを活用した生産管理，ビッグデータ活用した効率的実験デザイン

4. 社会インフラのスマート化 Smartization of social infrastructure

大量に配置したセンサーからのビッグデータの活用により，社会インフラのスマート化
（交通，電力供給，医療，都市、ビル管理システムなど）

ビッグデータが拓く世界（２） 日本学術会議提言「ビッグデータ時代に対応する人材の育成」

5. データに基づく意思決定・政策決定 Evidence based policy making

・経験と勘に代わるデータに基づく科学的意思決定・政策決定（マーケティング，品質管理，サプライチェーン効率化，リスク管理，公共投資や観光政策，環境対策）

・データ駆動型臨床試験，データ駆動型ジャーナリズム

6. 希少事象の発見とリスクの検知 Detection of rare event and risk

・稀な事象や隠れた関係性の発見（故障や災害の事前予測，列車等の運行安全の確保，金融リスク管理）

・ロングテール(稀少だが大きな価値を持つ事象）の発見によるイノベーション

・通信ログ解析（情報セキュリティの確保，不正発見）

・防災センサーとGPSデータ等の統合による災害リスクの早期検出（崖崩れ，地盤変化，地殻変動，火山噴火等）

7. 災害時オンライン対応 Real-time measure in case of disaster

モバイル情報やカーナビ情報などの位置情報を，個人情報保護の観点も考慮しつつ緊急時に避難，救助，ロジスティックスなどの支援に活用する

ビッグデータの背景： 社会・産業分野

Background of big data: In the society and industry

様々な変化 Various changes

- 「ものづくり」から「サービス提供」へ
- 「大量生産・大量消費」から「個別ニーズ対応」へ
- 「(普遍的)知識の応用」から「価値の創出」へ
- 「経験と勘」から「根拠に基づく意思決定」へ
  （政策決定 EBP，医療 EBM，精密農業）

Precision Medicine Initiative(2015)

https://obamawhitehouse.archives.gov/sites/default/files/longpage/wh_precision_medicine_header2.jpg

13

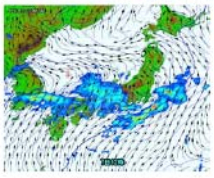
社会における科学の役割 Role of Science in the Society

専門家の経験と勘
Expert's experience and intuition

⇒

科学的方法
Scientific methods

- 占星術，航海術，錬金術
- 工業生産過程（ものづくり）
- 天気予報，経済予測
- マネジメント，マーケティング
- リスク管理，ファイナンス
- 科学的発見（発見科学）
- サービス
- 政策決定
- 研究コーディネーション



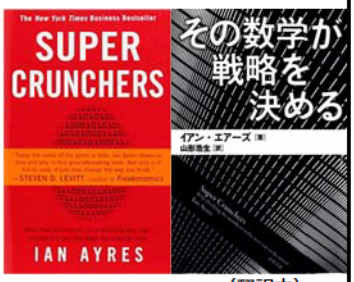
14

専門家 vs. データ分析 Expert vs. Data analysis

「大量データ分析」が「専門家の経験と勘」を凌駕する
"Big data analysis" surpasses "experience and intuition of experts"

Super Crunchers, Ian Ayers
Why thinking-by-numbers is the new way to be smart
翻訳版：『その数学が戦略を決める』（イアン・エアーズ著、文春文庫）

- ワインのヴィンテージ評価
- 野球のリクルーティング
- 人事採用
- カジノの顧客対応
- 航空会社顧客サービス
- 保険料の設定
- ネット販売の個別価格設定
- EBM　医療診断支援
- 判決予測、取引業者評価



（翻訳本）

15

歴史の転換点 The Transformation

歴史にも境界がある。

・・・数百年に一度、際立った転換が起こる。社会は数十年をかけて、次の新しい時代のために準備する。世界観を変え、価値観を変える。社会構造を変え、政治構造を変える。技術と芸術を変え、機関を変える。やがて50年後には、新しい世界が生まれる。

我々は今，まさにそのような転換の真っ只中にある。

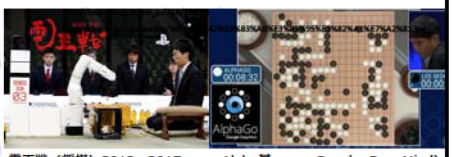
P.E. Drucker (1993)『ポスト資本主義社会』
Post-Capitalist Society

Every few hundred years in Western history there occurs a sharp transformation. We cross what in an earlier book I called a "divide." Within a few short decades, society rearranges itself – its worldview; its basic values; its social and political structure; its arts; its key institutions. Fifty years later, there is a new world. ..... We are currently living through just such a transformation.

歴史的局面の再現 Reappearance of Historical Moment

Baltimore and Ohio railroad 1830/12/25

佐野正博「技術の生存競争 ―「動力」に見る進化論」『週刊朝日百科 世界の歴史』第110号, 1991,p.693



電王戦（将棋）2012－2017 Alpha碁 << Google DeepMind)

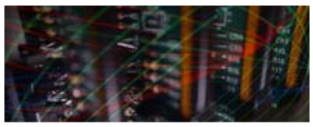
産業革命 Industrial Revolution
(18～20世紀)

蒸気機関による工業化社会の実現

情報通信革命 IC revolution
(21世紀)

ビッグデータ解析・人工知能による超スマート社会の実現

結果 人間と家畜を肉体労働から解放
Release humans and livestock from manual labor



人間を単純事務作業から解放
Release human beings from simple office work

17

「大半の人が思っているより遙かに早く変化は起きる」

Changes will occur much faster than most people think Yahoo CSO 安宅和人

https://commons.wikimedia.org/wiki/File:easterParade1900.jpg

T型フォード 1908

https://commons.wikimedia.org/wiki/File:Ave_5_NY_2_fl.bus.jpg

18

研究スタイルの変化 Changes of Research Style

「**認識の科学**」から「**設計の科学**」へ
From science for recognition to science for design

- **対象の変化** Change of object
  物理モデルだけでは解決できない複雑な対象・課題が重要に
  物理科学 ➡ 生命科学 ➡ 人間・社会・環境 ➡ CPS
- **目的の変化** Change of objective
  「真理の探究」から「予測、シミュレーション、知識創出、意思決定（制御、管理）」へ
- **モデルの変化** Change of model
  「物理（第1原理）**モデル**」から「目的達成のための**モデリング**」

CPS = Cyber Physical System (Society)

19

データ利用の変化： Changes in the role of data

仮説検証型から知識発展型へ From testing hypothesis to knowledge development

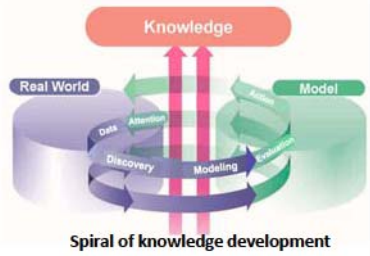
仮説検証型 Testing hypothesis

仮説 ➡ 実験 観測 調査 ➡ 検証

知識発展型 knowledge development

Data Data

知識 (Knowledge) ➡ Model ➡ 知識 (Knowledge) ➡ Model ➡ 知識



Spiral of knowledge development

ビッグデータ解析の時代へ Era of big data analysis

- 特定の目的のために精密に設計して取得したデータの解析
  ⬇ ・高精度（バイアス，分散），均一，少数データ
- 目的外のデータ，モニタリングデータ等のあらゆるデータを活用した解析
  ・大量・大規模・ヘテロ（形式，精度）

20

何をすべきか？ What should we do?

大波を活かすために・翻弄されないために To make use of big waves, to not be at the mercy of big waves

http://www.lairdhamilton.com/wp-content/uploads/2013/06/LairdBigWave2.jpg

21

ビッグデータ活用における課題 Nature and challenges of big data

- ビッグデータには膨大な知識や価値が埋もれている。
- 有効活用には課題がある。
  - 多くは構造化されていない
  - 大規模・価値密度が低い
  - 不均一（形式，精度，観測頻度，非定常性），スパース
  - 横断的検索やモデリングが困難

（1）共有化（公開，計算機可読，標準化）

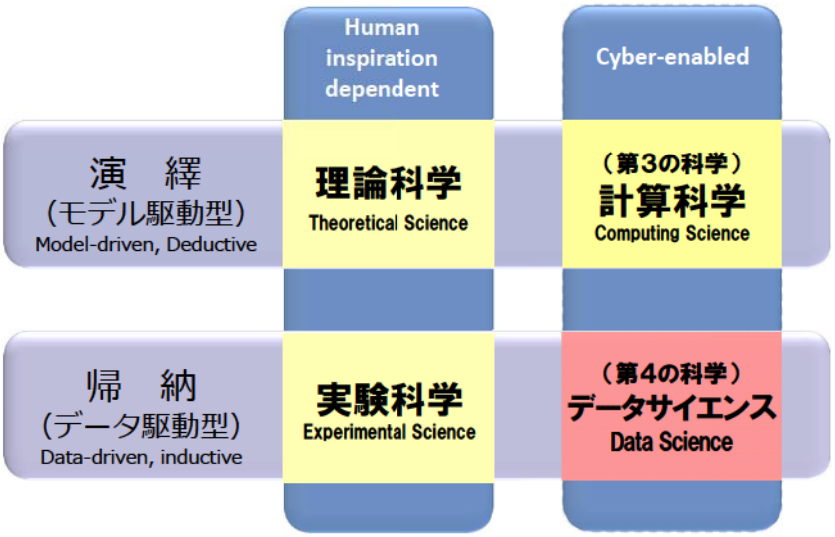
（2）共有化実現後の課題

ビッグデータを活用し，知識発見や価値創造を行う方法。

▶ ビッグデータ時代の新たな科学的方法論が必要

22

第4の科学 データサイエンス Fourth Paradigm: Data Science



23

データサイエンス Data Science

データに付加価値をつける技術 Technology to bring added value to data

支えるICTインフラ ICT infrastructure supporting data science

- ICTのコモディティ化
- IoTの進展（アナログプロセスのデジタル化）
- 帰納的推論技術の発展（機械学習、AI）
- 力技（ブルートフォース）の許容と成果
- 情報基盤
- 専門人材とデータリテラシー

関連分野： Related fields

統計科学（ベイズによる情報統合），パターン認識，機械学習（機械学習，深層学習，スパースモデリング），データマイニング，可視化
計算機科学，情報工学，データベース工学，

応用先: Applications

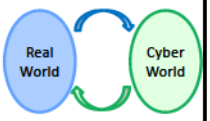
生物学，医学，工学，社会学，人文科学

24

海外の動向（米国）

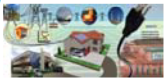
NSF（数学）

- 数理科学の重要課題「mathematical and statistical challenges posed by large data sets」(2004年度−)



情報学

- CDI (Cyber−enabled Discovery and Innovation, 2007−2011年度)
- CPS (Cyber−Physical Systems, 2009年度−)



- Materials Genome Initiative（2011年度−）
  計算ツール，実験ツール，数値データを材料イノベーション基盤とし，実験データやノウハウを蓄積したビッグデータを利活用するプロジェクト
- **Big Data Research and Development Initiative**
  - NSF, NIH, DOD, DARPA, DOE, USGSを通して2億ドルの財政支援により，**ビッグデータの最新技術を構築**.
  - スパコンとインターネットが過去の連邦政府の投資によって飛躍的に発展したと同様に，科学的発見，環境・生命医学研究，教育および国家安全保障における**ビッグデータ活用が飛躍的に進む**と指摘

- **欧州**
  - e−サイエンス（1999−） Grid, e−Science Center、FP7（2012−2014） Big Data Public Private Forum
  - Horizon2020 (2014-2020) Big Data

海外のデータサイエンス教育プログラム

185（4/2014） → 279 (7/2015) → 505（2/2016） → 515(5/2017)

2015年7月

	US	GB	IE	FR	NL	ES	Others	Total
Bachelors	15	2	0	1	0	0	2	**20**
Masters	126	37	7	5	6	5	28	**214**
Doctorate	11	1	0	0	0	0	2	**14**
Certificate	29	0	1	0	0	0	1	**31**
Total	**181**	**40**	**8**	**6**	**6**	**5**	**33**	**279**

2017年5月

	US	GB	IE	FR	NL	ES	Others	Total
Bachelors	37	5	1	1	0	1	4	**49**
Masters	281	40	7	6	7	8	41	**390**
Doctorate	16	1	0	0	0	0	2	**19**
Certificate	91	0	1	0	0	0	1	**93**
Total	**425**	**46**	**9**	**7**	**7**	**9**	**48**	**551**

アメリカの大学における統計教育

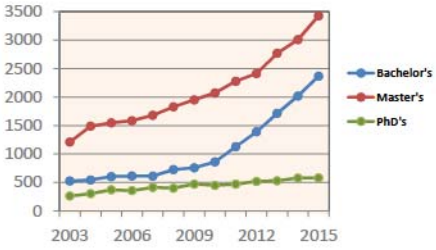
● アメリカの動向

2015年　上級コース試験(AP)受験者　19.6万人(2009年11.7万人、2005年7.7万人)
2010年　Statistician　28,000人　US Bureau of Labor Statistics(BLS)
2015年　86,000人

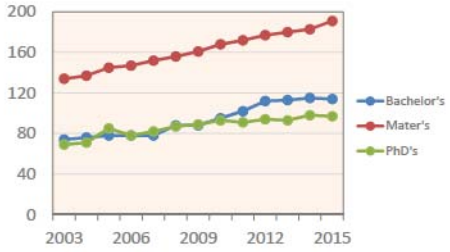
ビッグデータ活用に必要な人材　14万人～18万人不足(2018年)　McKinsey Global Institute

統計学士取得数(78%増　2011/2003, 48%増　2011/2009)
統計関連博士(数理統計, 生物統計, 計量経済, 経営統計)取得数(632人(2002), 841人(2007), 991人(2012)

学位授与者数（2003年－2015年）



学位授与大学数（2003年－2015年）



データサイエンティスト育成プログラム

海外のBootcamps 27（US 19，UK 2，その他 6）

● **Insight Data Science Fellows Program**

- ✓ ６週間のデータサイエンティスト養成プログラム
- ✓ IT，流通系30社連携
- ✓ 逆インターンシップ，Open dataによる実習
- ✓ Hard Scientistsのビッグデータ科学再教育
- ✓ 100%の就職率（平均初任給約1000万円）
- ✓ NYでも開始

INSIGHT

Google　twitter　facebook
Microsoft　LinkedIn　Square

● **NYC Data Science Academy**

- ✓ データサイエンティスト育成。2013年11月創設。
- ✓ Bootcamps, Corporate Training, Part-time Coursesなどのコースがある
  - "Bootcamp"の授業料は$16K（200万円）/12週とかなり高額
  - 10週間の授業のあと、現実の問題を解決するケーススタディを行いプログラムを書くことを義務付け
- ✓ 企業からの求人は非常に多い。
- ✓ Kaggleでもいつも上位の成績を収めている.
- ✓ 北米以外では、台湾、韓国、中国、南アでも展開。

NYC DATA SCIENCE ACADEMY

海外のデータサイエンス研究組織

2014年 9月　6研究組織 (NIH, Virginia, Rochester, UCB, UC London, Columbia)
2014年11月　9研究組織
2016年 2月　25研究組織

アメリカ：
ニューヨーク大学：　Center for Data Science
UC Berkeley；　Berkeley Institute for Data Science
コロンビア大学：　Data Science Institute
スタンフォード大学：　Stanford Data Science Initiative
CALTech：　Center for Data Driven Discovery
Johns Hopkins大学：　Institute for Data Intensive Engineering and Science
その他 Michigan, Rochester, Virginia, N. Carolina S., Syracuse等計17大学

英国：
Imp. College London　Data Science Institute
Oxford:　Emirates Data Science Lab
Cambridge等計4大学

オーストラリア, ドイツ, オランダ, シンガポール

アメリカのデータサイエンス推進組織

- **NSF BD Hubs (Big Data Regional Innovation Hubs)**
  - ✓ Obama Big Data Initiativeの一環
  - ✓ 研究ではなく，ハブ構築の支援（11/2015公表）各$4~5M
  - ✓ 全米４地域にハブを構築
    - 北東部（コロンビア大学：エネルギー，金融，環境，DS育成）
    - 中西部（イリノイ大学：　農業，スマートシティ）
    - 南部（ジョージア工科大，Nカロライナ大：健康，産業BD）
    - 西部（UCSD,UCB,UW：資源保護，医薬）
  - ✓ NSF BD Spokes
    - 各地域のHubを中心にSpokesを構成（大学，企業，政府，NPOなど）
    - 各スポークは特定の問題にフォーカス
    - 2/25/2016（応募締切）各スポーク$1M(3年間)

- **DSE（Data Science Environment：Moore財団，Sloan財団）**
  - ✓ Moore財団
    - ４分野（環境，サイエンス，医療，SFベイエリア：＄70M程度）
  - ✓ Sloan財団
    - 科学技術の研究・教育，経済，バイオテロ対策（$75-80M）
  - ✓ 約２年をかけて支援対象を選定（15→3）
  - ✓ UW, UCB, NYUを支援

国内におけるビッグデータ活用の動き Trends in Japan

- 情報・システム研究機構設立 (2004)
  統計数理研究所，国立情報学研究所，国立遺伝学研究所，国立極地研究所
- 情報爆発プロジェクト（2006－2010　科研費特定領域研究）
  情報爆発時代に向けた新しいIT基盤技術の研究
- 情報大航海プロジェクト（2007-2009　経済産業省）
- JST さきがけ「知の創生と情報社会」（2008－）
- JST Crest・さきがけ（2013－）
  「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」
  「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化」
- AIP（2016－ 人工知能/ビッグデータ/IoT/サイバーセキュリティ統合プロジェクト）
- JST CREST・さきがけ（2016－）
  「イノベーション創発に資する人工知能基盤技術の創出と統合化」
  「計測技術と高度情報処理の融合によるインテリジェント計測・解析手法の開発と応用」
- MI²I（2016－ 情報統合型物質・材料開発イニシアティブ）
- 「数理及びデータサイエンスに係る教育強化」６拠点校（2016－）

31

日本の統計教育の特殊事情 Special circumstances of statistical education in Japan

我が国の統計人材育成方式

統計学科，統計学専攻を設置せず，応用分野（医学，工学，農学，数学，経済学，心理学，社会学など）で専門的人材を育成してきた（分野点在方式）．

○ 現実の問題に根差した研究・教育の実現（先人の工夫）

- 当該分野の課題に特化した方法論になりがち．抽象化の不足
- 新分野開拓，他分野への転向の困難
- コミュニティ形成不足：日本統計学会 1500人，統計関連学会連合 3000人?
  （米ASA 18000人，英RSS 7200人）

Deep Analytical Talent (米24730人，中17410人，英8340人，日3400人 MBI レポート)

日本学術会議勧告・提言 Recommendation and proposal by Science Council of Japan

「統計学の大学院研究教育体制の改善について（勧告）」（1983年11月）
- 大学院専攻（教員定員18名，学生定員30＋6名程度）の複数設置
- 実現は総合研究大学院大学（統数研，1988年設置，博士5名）のみ

「ビッグデータ時代における統計教育・研究の推進について(提言)」(2014年8月)

産学官懇談会提言 Recommendation by Industry-government-academic conference

「ビッグデータの利活用のための専門人材育成について」（2015年7月）

32

ビッグデータ時代に対応する人材の育成

"Fostering Data Scientists for Big Data Era", Science Council of Japan

日本学術会議 情報学委員会
E-サイエンス・データ中心科学分科会（2014年9月11日）

1．はじめに
2．海外の動向
3．我が国の現状と人材育成に関する課題
4．ビッグデータ活用に必要な要素技術と人材育成
5．提言 Recommendations

- ●提言1　データ中心科学を専門とする教育組織の設置
  Establishment of an educational organization specializing in data-centered science
- ●提言2　基幹的研究組織内における恒久的なデータ解析部門の設置
  Establishment of a permanent data analysis departments within the core research organization
- ●提言3　日本版インサイト・プログラムの早急な設置
  Urgent installation of the Japanese Version Insight Program
- ●提言4　データサイエンティストの資格の制定
  Establishment of data scientist qualification

33

4-1. データ中心科学の要素技術

- **データ解析** **ビッグデータからの深い知識獲得のための方法**
  統計的モデリング，ベイズ推論，機械学習，データマイニング，テキスト検索，Web情報解析，自然言語処理，最適化
  - ・データ同化法
  - ・インピュテーション技術（内外挿，不完全データ・異常値処理）
  - ・新NP問題の解決
  - ・高次元空間の構造探索とモデル化
  - ・異種情報統合による個人化技術（サービスのテーラーメード化）
  - ・隠れた関係の検出，特異性の発見，因果推論の実現
- **データ可視化** **膨大な高次元データや計算結果を人間が把握できるようにするための技術**
  次元圧縮，特徴抽出，パターン認識や画像処理
- **データ処理技術** **大量の散在するデータを処理するための技術**
  分散処理，並列処理，HPC，ストリーミング計算，巨大データベース，リンケージ技術，クラウド計算などの情報処理技術

4-2. データサイエンティストの要件

データサイエンティストは問題の本質の把握，定式化，データ取得，分析，予測，知識獲得，意思決定，課題解決，評価の全過程に関与

要素技術

- **データ解析法**
- **データ可視化**
- **ビッグデータ処理技術**

データリテラシー

- 問題発掘、問題解決戦略立案能力
- データ収集能力，キュレーション能力（データの選択，前処理，クレンジング）
- データの背景を見抜き，関連するデータを見出す力
- データ分析結果の業務や事業への実装能力
- 異分野研究者・事業者との連携・コミュニケーション能力
- 研究倫理，個人情報保護

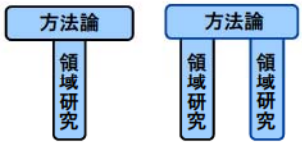
35

4－3. データサイエンティストの育成方法

● **データ中心科学と融合研究の推進に必要な人材**

- データ解析、可視化，データ処理，モデリング、知識発展の方法
- 領域科学の知識と理解
- 課題設定，企画立案能力
- コミュニケーション能力（異分野交流）
- 研究倫理，個人情報保護

⇨ T-型，Π-型研究者



● **データサイエンティストの育成方法**

（1）主専攻：データサイエンス（統計数理），副専攻：領域科学
（2）領域科学の博士取得者の統計・数理・情報再教育

5. 提言の内容

日本学術会議提言「ビッグデータ時代に対応する人材の育成」

●提言1 データ中心科学を専門とする教育組織の設置

データ中心科学を専門とする高等教育研究組織を設置し，ビッグデータ解析技術等の分野横断型の学問を主専攻、領域科学を副専攻として，T型・Π型人材を育成すべきである。

●提言2 基幹的研究組織内における恒久的なデータ解析部門の設置

データ解析が重要な役割を果たす研究領域の基幹研究所内に部署横断的なデータ解析専門の恒久的な研究部門を設置すべきである。

●提言3 日本版インサイト・プログラムの早急な設置

アカデミアの人材と産業界が要求する人材の乖離を埋めるために、既卒の博士研究員の横断型科学の習得，現実の課題への挑戦と異分野交流の経験による再教育を実施し、データサイエンティストを育成すべきである。

●提言4 データサイエンティストの資格の制定

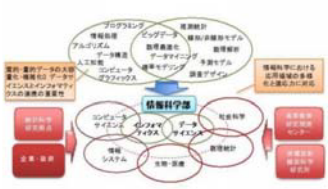
ビッグデータ時代の科学技術研究及び産業界のイノベーションを先導するトップタレントとしての質保証の観点からデータサイエンティスト資格の制定が望ましい。

37

データサイエンス学部等の新設

滋賀大学（2017）
データサイエンス学部新設

広島大学 情報科学部（2018予定）
・データサイエンスコース
・インフォーマティクスコース



横浜市立大学 データサイエンス学部（2018予定）
・データサイエンス学科

38

ビッグデータの利活用のための専門人材育成について

情報・システム研究機構
ビッグデータの利活用に係る専門人材育成に向けた
産学官懇談会　（2015年7月30日）

1．データサイエンティスト育成の必要性と我が国の課題
2．我が国におけるデータサイエンティストへの要請
3．データサイエンティスト人事育成のあるべき姿と実現に向けた仮説
4．具体的施策
5．まとめ

- 提言1　500名規模の「棟梁レベル」人材育成とトリクルダウン
- 提言2　主要10大学程度での人材育成による大学教育加速
- 提言3　全学的教養教育の実施，国家レベルフラグシッププロジェクトの推進

39

ビッグデータの利活用に係る人材育成に向けた産学官懇談会

座長	北川 源四郎	情報・システム研究機構 機構長
委員	安宅 和人	ヤフーCSO ／データサイエンティスト協会 理事
	榎本　剛	文部科学省 研究振興局 参事官(情報担当)
	岡本 青史	富士通研究所
	北山 浩士	文部科学省 高等教育局 専門教育課 課長
	佐藤 俊哉	京都大学医学研究科 教授
	長谷川 眞理子	総合研究大学院大学 理事・副学長(教育担当)
	樋口 知之	統計数理研究所 所長
	丸山 宏	統計数理研究所 教授 データサイエンティスト育成ネットワーク事業 実施担当責任者
	丸山 文宏	富士通研究所
	渡辺 美智子	慶應義塾大学健康マネジメント研究科 教授／(独)統計センター理事
オブザーバ	粟辻 康博	文部科学省 研究振興局 数学イノベーションユニット次長 基礎研究振興課 融合領域研究推進官
	金井 学	文部科学省 高等教育局 専門教育課 情報教育推進係長
	栗原 潔	文部科学省 研究振興局 参事官(情報担当)付専門官
	土生木 茂雄	文部科学省 高等教育局 専門教育課 視学官
	山路 尚武	文部科学省 高等教育局 専門教育課 課長補佐

産 Industry　学 Academia　官 Government　ROIS

40

データサイエンティストへの要請（産業界からの要請）

Requests for data scientists (from industry/Business)

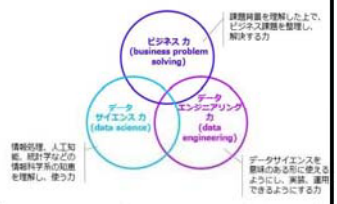
●データサイエンティストに求められるスキルセット

① データサイエンス力（Data science）
情報処理，人工知能，統計学などの情報科学系の知恵を理解し，使う力

② データエンジニアリング力（Data engineering）
データサイエンスを意味のある形に使えるようにし，実装，運用できるようにする力

③ ビジネス力（Business problem solving）
課題背景を理解した上で，ビジネス課題を整理し，解決する力

データサイエンティストに求められるスキルセット



（データサイエンティスト協会スキルセット定義委員会報告書より転載）

●データサイエンティストに求められる能力

① 顧客の課題をデータ分析や情報技術に落としこむ翻訳力
② 課題領域を見通して本質的な問題を見抜く能力
③ 課題解決のための各分野のエキスパートを動員できる能力

●産業界からの要請

① 大学でのデータサイエンティスト育成のためのプロフェッショナル教育
② 中等・高等教育を含む理系素養・データリテラシーのテコ入れ
③ 国家レベルのビッグデータ活用フラグシップ・プロジェクト

41

データサイエンティストへの要請（アカデミア・地方自治体）

アカデミアからの要請

●あらゆる分野の研究者は，同時にデータサイエンティストでなければならない
●トップレベルのビッグデータ分析手法を研究し，第４の科学を牽引できるトップレベルの研究者が必要
●リテラシーレベルでは，データに基づく思考ができるようにする
（参考）京都大学の「統計入門」
●分野を繋ぐ本質的な媒介者（T型・Π型人材）の育成
データサイエンスは複数の分野の研究者が視野を共有するための共通言語
（参考）統数研の「統計思考院」

地方自治体からの要請

●オープンデータの動き，エビデンスに基づく効果的施策立案・評価を担う人材確保が急務
●データを戦略的に活用する部署の設置が必要
●全国レベルの底上げのために，全国的に一定数の人材育成拠点が必要

42

データサイエンス人材のレベル

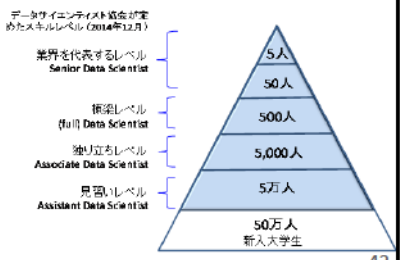
データリテラシー
理系・文系に限らずすべての学生が持つべき素養．統計的概念，データに基づく思考や問題解決の基礎概念，ITリテラシー，データに関する研究倫理（全大学入学者を想定）

見習い（基礎能力）レベル
すべてのデータサイエンティスト(研究者・実務家)およびマネジメントが持つべき共通能力．理系の修士は全て，文系でも経済・経営，心理学，言語学等データ関連の分野．適当な指導の下で，ビッグデータ活用プロジェクトの一部を担当できる．（全理系修士入学者を想定）

独り立ちレベル
専門能力（ビジネス，データサイエンス，データエンジニアリング）のいずれかの専門的能力を持ち，自らのイニシアティブで高度なデータ分析，問題解決ができる．修士・博士での具体的なPBLの経験．（資本金10億円以上の全企業（6000社）を想定）

棟梁レベル
DSのチームを率いて，組織におけるビッグデータ利活用を先導できる人．実務や大学院社会人コース等で育成（想定数は独り立ちレベルの1/10）

指導的データサイエンティスト
学界： DSの最先端を切り開くワールドクラスの研究者
産業界：業界におけるビッグデータに基づくイノベーションを牽引できる人



43

人材育成の具体的施策：リテラシーの醸成（50万人規模）

- **高校教育・大学教養の講義で、データとその利活用で世界が大きく変わっている重大性を教える。**たとえば、
  - 高校生・大学生がワクワクするような啓発書・教科書をつくる。
  - 社会でどう使われ役立っているかを示す実例のビデオ素材をつくる。
  - 活躍中のDSをプールし、大学や高校に適宜派遣し講義を担当してもらう。
- **大学基礎教育にデータサイエンスを取り入れる。**
  - 大学124単位の内、共通教育で例えば4単位、専門教育では専門に応じて例えば2から6単位をデータサイエンス（統計）に割り当てる。この際、核となる週1時間の講義にコンピュータ実習や問題を解く演習もセットにする。
  - これに合わせて、基礎統計教育も見直す。
- **社会一般の興味を惹くための施策を実施する。**
  - 全日本的なデータサイエンスに関するコンテストを実施。スーパーグローバルハイスクール指定校等の先進的な取り組みを行っている高校にも積極的参加を呼びかける。
  - gaccoにおける「社会人のためのデータサイエンス入門」のように、MOOCを利用した一般向け教材を充実させる。

44

人材育成の具体的施策：見習いレベルの育成（5万人規模）

- データサイエンス力の基本的スキルとして、統計学、機械学習、最適化、プログラミングやデータ可視化の基礎を学習させる。学部・大学院におけるデータサイエンスの参照基準を早急に作成する。
- 大学院に、ジョイントディグリーを導入し、専門科目と共にデータサイエンスを副専攻などの形で学べるようにする。逆に、データサイエンスを主専攻とする学生が、データサイエンスの適用分野を副専攻として学べる機会も提供する。
- スケールアウトの方策として、MOOC を積極的に利用する。
  - データサイエンス向けのコンテンツの充実。
  - MOOC修了者に対する修了書の発行、大学における単位認定などにより、MOOCの位置付けを社会的に認知させる。
- 社会人に対しては、ミドルマネジメント層を含む広い対象に、再教育のプログラムを提供し，各大学に展開する。社会人の学び直しを推進するため，「職業実践力育成プログラム」認定制度の活用も考えられる。

45

人材育成の具体的施策： 独り立ちレベルの育成（5000人規模）

- 問題設定能力、問題解決のための戦略立案能力、データの収集・キュレーション能力、データ分析結果の業務や事業への実装能力、異分野研究者や事業者との連携・コミュニケーション能力、研究倫理、情報セキュリティの能力を備え、独立してデータサイエンスを推進できるレベルを目指す。
  - データ解析スキル：MCMC法、データ同化、インピュテーション技術、高次元空間の構造探索とモデル化、異種情報統合による個人化技術、隠れた関係の検出、特異性の発見、因果推論など
  - データ可視化技術：次元圧縮、特徴抽出、パターン認識など
- 大学院においてPBLに基づく専門育成プログラムを推進する。
  - 「分野・地域を越えた実践的情報教育協働ネットワーク(enPiT)」、「先導的ITスペシャリスト育成推進プログラム」などにおけるPBLの取組が参考になる。
- 企業との連携により、企業の事例のPBL化、インターンシップによって実務機会の提供。
  - 産学連携では経団連主体のCeFIL（高度情報通信人材育成支援センター）が参考。

46

人材育成の具体的施策： 棟梁レベルの育成（500人規模）

- データサイエンスの最新の手法群と、新しい応用分野に精通している必要がある。このため、ビッグデータ・データサイエンスの最先端の手法・応用の研究・開発及びそれらに精通した人材育成を行う国家的な拠点を設置する必要がある。
- データサイエンティストとしての実務経験のある社会人を、棟梁レベルに育成する集中的プログラムを上記拠点で開講する。
  - 最先端の手法をPBLにより実地で経験し、各応用領域での最新の成果をケーススタディとして学ぶ.
  - 棟梁レベルデータサイエンティスト間の人的ネットワークを形成する。
  - 社会人を受け入れてデータサイエンスにおける経験と人脈を形成させる。
- 科学の諸分野の博士号取得者等を産業界やアカデミアにおいて活躍できるように再教育する。
  - 米国のInsight Data Science Fellow Programが参考になる
  - 統計数理研究所「統計思考院」のDS育成プログラムを量的に拡充する。

47

人材育成の具体的施策:指導的データサイエンティスト（50人規模）

- 指導的データサイエンティストは、系統的に育成するのは難しい。DS拠点において、世界最先端の手法・応用の研究・開発を推進することによって、このような指導的データサイエンティストが生まれてくる土壌を醸成することが大切である。
  - 国レベルで、ビッグデータのフラッグシップ・プロジェクトを実施し、その中で指導的データサイエンティストが活躍できる場を提供する。
- 同時に、このような才能のポテンシャルを持つ者を若いうちに、発掘し、十分な機会を与える必要がある。このための施策として、以下の2点を提案する。
  - 定期的にデータサイエンス・ハッカソンを実施。
  - 産業界で活躍する指導的DS人材の育成のために、才能のある若い者にメンターをつけ、資金等の援助を与えると共に人脈を形成する機会を提供する。（参考：IPA「未踏IT人材発掘・育成事業」やJST「さきがけ」）

48

「中抜き」仮説とその対応

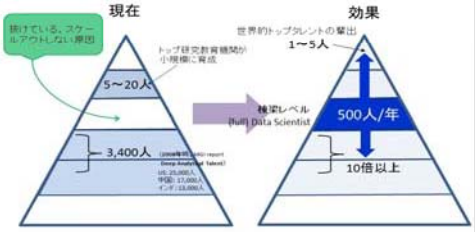
「中抜き」仮説

我が国に於ける最も深刻な問題は「棟梁レベル」の人材が育っていないこと

棟梁レベル：
チームを率いて，組織におけるビッグデータ利活用を先導できる能力を持った人．

原因

棟梁レベルの人材が活躍できる場がなかった



データサイエンス人材育成の鍵は棟梁レベルの育成。

- 棟梁レベルの人材が年500人規模で育てば、その中から世界的なトップタレントが現れてこの分野全体を引っ張っていくことが期待できる。
- また同時にこの棟梁レベル人材が人材育成を支援することで、独り立ちレベル以下の人材育成が促進され、スケールアウトが進むことを期待することができる。（トリクルダウン効果）

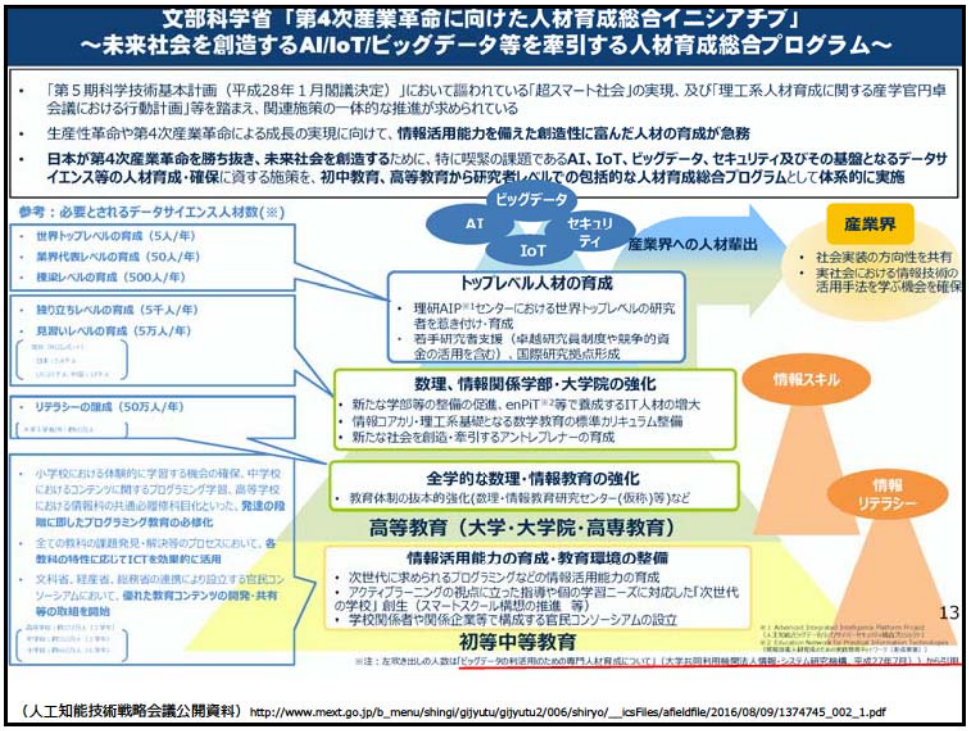
49

提言

1. 我が国の問題の根源は、棟梁レベルの決定的不足にある。この解決のために国家レベルの拠点を設置して、年500名規模の**「棟梁レベル」**の人材育成をめざし、上層への成長や下層へのトリクルダウン効果も狙う。
2. リテラシーレベルや独り立ちレベルの大学教育を加速させるために、主要10大学程度で本報告書の提案に基づく人材育成をスタートすると共に、MOOCなどのオンライン教材を整備し、全国への波及効果を狙う。
3. 社会全体のリテラシーやアウェアネスを向上させるために、全学的教養教育の実施、国家レベルのフラッグシップ・プロジェクトの推進、コンテストの開催、映像素材の充実などの取組を行う。

これらの方策を実現するにあたっては、データサイエンスを副専攻とするジョイント・ディグリー制人材育成の推進やスキル認定制度も有効と考えられる。

ビッグデータの利活用のための専門人材育成について　50



（人工知能技術戦略会議公開資料）http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu2/006/shiryo/__icsFiles/afieldfile/2016/08/09/1374745_002_1.pdf

数理・データサイエンス教育強化方策(文科省)

主要6大学に数理・データサイエンス教育研究センター（仮称）の整備

●センターのミッション
- 数理・データサイエンスの全学的教育実施，価値創出ができる人材育成
- コンソーシアム形成
  - 標準的カリキュラム・教材を協働して作成
  - 取組成果の他大学への展開・波及
  - 大学，産業界，研究機関等と連携したネットワーク形成

●実施体制
- 専任教員の配置
- 幅広い分野の教員の参画

●教育内容
- 数理的思考とデータ分析・活用能力をバランスよく修得
- 戦略的設定（課題の発見，データの取得・一次処理）を行ったうえでの数理的モデリングによる価値創出

●実践教育の在り方
- 他分野・他大学・他機関との交流の場の設定
- 数理・データサイエンス × 多分野・産業プログラムの実施
- 標準カリキュラム，オープンデータセット，ケーススタディの活用

●拠点校
北大，東大，滋賀大，京大，阪大，九大

52

計測と高度情報処理の融合

CREST・さきがけ複合領域
「計測技術と高度情報処理の融合による
インテリジェント計測・解析手法の開発と応用」
募集説明会

総括（CREST担当） 雨宮 慶幸
東京大学 大学院新領域創成科学研究科 教授
副総括（さきがけ担当）北川 源四郎
情報・システム研究機構 機構長

科学技術振興機構

Japan Science and Technology Agency

本領域のねらい

- 多様な計測技術に最先端の情報科学・統計数理の研究を高度に融合させることによって，**インテリジェント計測・解析手法**を開発
  - これまでは捉えられなかった物理量・物質状態やその変化等の検出
  - これまでは困難であった測定対象が実際に動作・機能している条件下でのリアルタイム計測等を実現
  - 直接観測できない内部構造や物理的に存在しない潜在要因等の推定
- 方法的・基盤的研究
  - ベイズ推論，データ同化，スパースモデリング，機械学習，最適化技術，画像解析，信号処理等の広範な逆解析技術を中心にした情報科学・統計数理による計測対象の特徴量解析手法や大量データの迅速・高精度解析手法等の開発
- 具体的な計測課題への適用
  - 物質・材料，生命・医療・創薬，資源・エネルギー，地球・宇宙，人間社会，Web空間等，科学技術全般における新現象の発見，原理の解明，新たな知識獲得，意思決定等
- 数理情報計測科学の確立

Japan Science and Technology Agency 0

応募してほしい課題イメージ（計測対象の観点から）

- 間接計測データ（宇宙，惑星，リモートS）
- 直接計測データ（微量隕石，アイスコア）
- 超大量・大規模データ（大型大気レーダー）
- 超微量計測データ（分子計測）
- バイオイメージング（脳活動計測，神経系トラッキング）
- 内部構造推定（CT, fMRI, 地下構造），ネットワーク推定
- 隠れ因子（トレンド，季節成分，景気動向，コホート効果）

Japan Science and Technology Agency 2

解析の観点からの課題イメージ（例示）

－ 計測の限界突破に必要な情報・解析技術 －

- 逆解析技術（データ同化，逆問題，ベイズモデリング）
- 情報抽出技術（正則化，スパースモデリング，機械学習）
- 情報統合技術（ベイズモデリング，データ同化，個別化技術）
- 最適化技術（凸最適化，組合せ最適化）
- 高度データ処理（画像解析，信号処理）
- 高次元時空間系列解析（状態空間モデリング，予測）
- 計測限界突破技術（超微量，超大規模，動態）

Japan Science and Technology Agency 53

情報統合型物質・材料開発イニシアティブ（NIMS-MI²I）

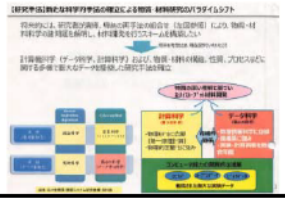
蓄電池材料G
1. 全個体電池
2. 多価イオン電池

磁気・スピントロニクスG
1. 強力永久磁石
2. 磁気メモリ

電熱制御・熱電材料G
1. 高・低熱伝導材料
2. 熱電材料

Data Science Group
Modeling Group
情報統合型物質・材料開発のための，方法論の開発，ツールの開発・整備

Data Platform Group
情報統合型物質・材料開発支援システムの構築



NIMS 外部諮問委員会資料から編集

54

理研AIPセンターの具体的研究内容

- 世界をリードする革新的人工知能基盤技術を構築する
- サイエンスや実社会などの幅広い"出口"に向けた応用研究を進める
- 未来の科学研究に必要となるデータ構造、データ取得技術のデファクト・スタンダード化、世界標準化を図る

＜運営上考慮する点＞
- 戦略的な研究開発ポートフォリオを構築
- 外国人研究者比率30%以上を目指す
- 研究者の長期的・安定的な雇用を確保
- 研究者が世界を飛び回れるような自由度を確保
- NICT・産総研や大学等とアンダーワンルーフでの連携・協力を実現

他機関と連携

超高齢化社会へ向けた医療サポート
- 動画認識・センサ情報解析・アクチュエータ制御技術の融合による複雑な臓器のような超柔軟・変形物体の認識とマニピュレーションの実現
- 多様な生体センサ・電子医療記録・医療画像情報のマルチモーダルな情報の自動的体系化と予後予測
- 対話・音声認識・自然言語処理技術と認知モデル・行動解析の融合による、高齢者の認知機能の維持向上の実現

老朽化が進むインフラへの対応
- 橋梁などの寿命について、危険箇所の特定と自律的な学習を機器の完全自動検査により行い対処
- 検知の機械学習技術と、画像やセンサのマルチモーダルな情報解析により、過去に例のない危険個所の自動推定

AIPセンターで実施する中核的な研究開発

革新的アルゴリズムに基づく基盤研究開発
1. 深層学習等のビッグデータを用いた学習のさらなる抜本的な革新：現在の技術が適用できないより複雑な構造を持ったデータに対応するための、深層学習等の現在主流の機械学習手法を深化させる革新的超高速学習アルゴリズムを開発する。
2. スパース（疎）・不完全・超高次元のデータからの高度で高精度な学習の実現：実社会の実データへの応用で重要となる、偏った情報・不完全な情報・超高次元の情報から高度で高精度な学習が可能な革新的不完全情報学習アルゴリズムを開発する。
3. 実環境におけるストリーミングデータからの頑健なリアルタイム学習の実現：多様なセンサーが大規模に用いられる次世代社会で重要となる、次々と与えられるデータを即座に学習し・結果をフィードバックできる革新的リアルタイム学習アルゴリズムを開発する。

人工知能技術の統合化研究開発
1. 学習のための学習：　学習アルゴリズムの選択・調整パラメータのチューニングを自動化。
2. ハードウェアへの最適化：　CPU GPUだけでなくDisk I/O等も考慮した新しいデータ処理パラダイムを開拓。CPU GPU等ハードウェア構成も考慮した超高速並列探索技術の開発。
3. データ収集の最適化：能動学習・ベイズ最適化などの本質を解明して適用し、データ収集を最適化することで実装における効率化につなげる。

ブライバシーに関する研究
- 医師・弁護士・政治家等専門家の判断にAIが介在する場合等の倫理的な問題の解決
- 深層学習等の学習アルゴリズムによる予測・判断について、プライバシー等を保証するための基盤技術の構築
- AIにおけるプライバシーの適正制御技術や差別発言を未然に防ぐ等のための技術の開発

他機関と連携

甚大な自然災害への対応
- ビッグデータ解析技術とシミュレーション技術の統合により、甚大な災害とそれによる影響を複雑な社会状況も含めて高精度に予測
- 予測に加えて被害の最大限の抑制・迅速な復旧の促進を含む「対応」までも対象とした自律学習機能の獲得

細分化が進む科学研究への対応
- 論文や実験結果等の膨大な重複検定等によりこれまで埋もれていた新たな科学発見
- 能動学習・ベイズ最適化に基づいて、実験計画を策定
- 自然言語処理を駆使し、蓄積型情報を構造解析し、フロー情報に基づき最新の研究開発動向や新たな研究のヒントを示すスマート・アラートシステムの構築

（人工知能技術戦略会議公開資料）http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu2/006/shiryo/__icsFiles/afieldfile/2016/08/09/1374745_002_1.pdf　55

革新的な人工知能基盤研究開発の内容について

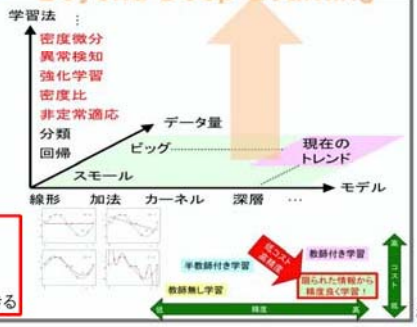
現状
- 欧米：巨大民間企業等が数百億～数兆円規模の莫大な予算を投じて研究開発を開始（Google, Microsoft, Facebook, Amazon, Toyota, OpenAI,…）
- 国内：政府が中心となって，数十億～数百億円規模の予算を幅広い分野に配分
- 様々な人工知能技術に関する国際会議では、限られた研究者しか活躍できていない。

課題
- ビッグデータ＋深層学習＝究極の人工知能ではない？
- 多数のデータを用いれば、古典的な最近傍分類等であっても必要な性能の発揮が可能。
- センサーの数を増やすと、データの次元数は増加 → いくらビッグデータを集めても、データはまばら
- ビッグデータの現実：同じさいころを多数回振れない、そもそも答えのないデータばかり、限られた情報からの学習が重要

10年先を目指した革新的な人工知能基盤研究開発

革新的な次世代の人工知能基盤技術を構築
- 教師付き学習：人間が教師となり、コンピュータを学習させる
  例：脳波によるコンピュータの操作に適用（独Fraunhofer研究所との共同研究）
- 強化学習：エージェントが試行錯誤を通じて学習する
  例：ヒューマノイドロボットの運動制御に適用（NICT/ATRとの共同研究）
- 教師なし学習：コンピュータが人間の手を介さずに学習する
  例：データの可視化、クラスタリング、密度比推定

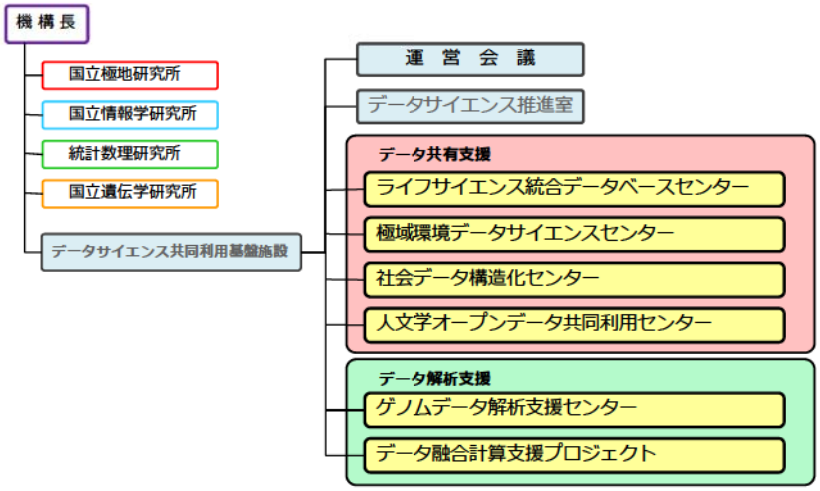
究極の人工知能技術となる汎化能力の実現
- 汎化能力：教わっていないことを、過去の事例から類推する能力
- まばらなデータの間を補完・予測できる
- ビッグデータを用いれば、真の答えがわからなくても、簡単に予測できる



（人工知能技術戦略会議公開資料）http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu2/006/shiryo/__icsFiles/afieldfile/2016/08/09/1374745_002_1.pdf　56

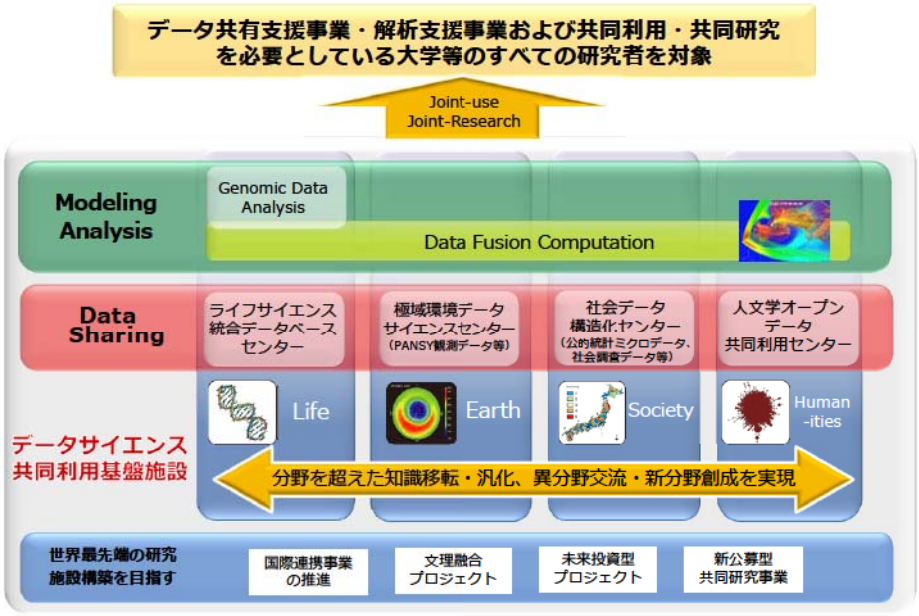
データサイエンス共同利用基盤施設

大学共同利用機関法人 情報・システム研究機構(ROIS)における**ビッグデータ活用のための基盤整備事業**



57

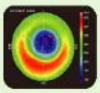
データサイエンス共同利用基盤施設



58

データ共有基盤: Data Sharing Platform

Life Science Data
Database Center for Life Science (DBCLS)



Earth Environmental Science Data
Polar Environmental Data Science Center

Social Science Data
Social Data Structuring Center

Humanities Data
Humanities Open Data Joint-Use Center

59

生命科学: ライフサイエンス統合データベースセンター(DBCLS)

ライフサイエンス・データベースの統合化を実現するための
研究開発を推進（JST/NBDCとの共同研究事業）

事業内容:

- 現存する 270のLSデータベースの統合
- DB統合化における国際標準化推進
- データベース統合化のための技術開発
- 日本のライフサイエンス・データベースの保存
- アノテーション
- ライフサイエンス・DBポータル サイト
- ライフサイエンス分野の出版情報
- キュレーターおよびアノテーターの育成
- BioHackathon: 開発者向けの国際ワークショップ
- 生命科学におけるオープンリサーチデータの共有・利用支援

統合化推進事業の方向性

Web サービスを利用したDB統合
計算機に適したデータアクセス

↓

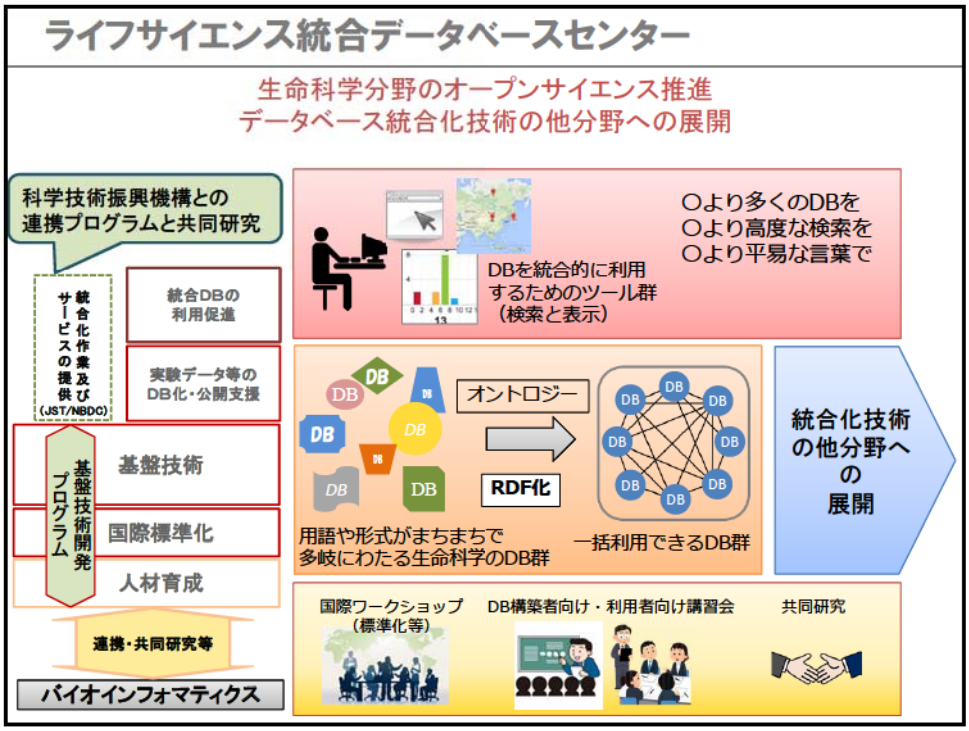
RDFデータによるDB統合
RDF=共通のデータ形式とグローバルなID

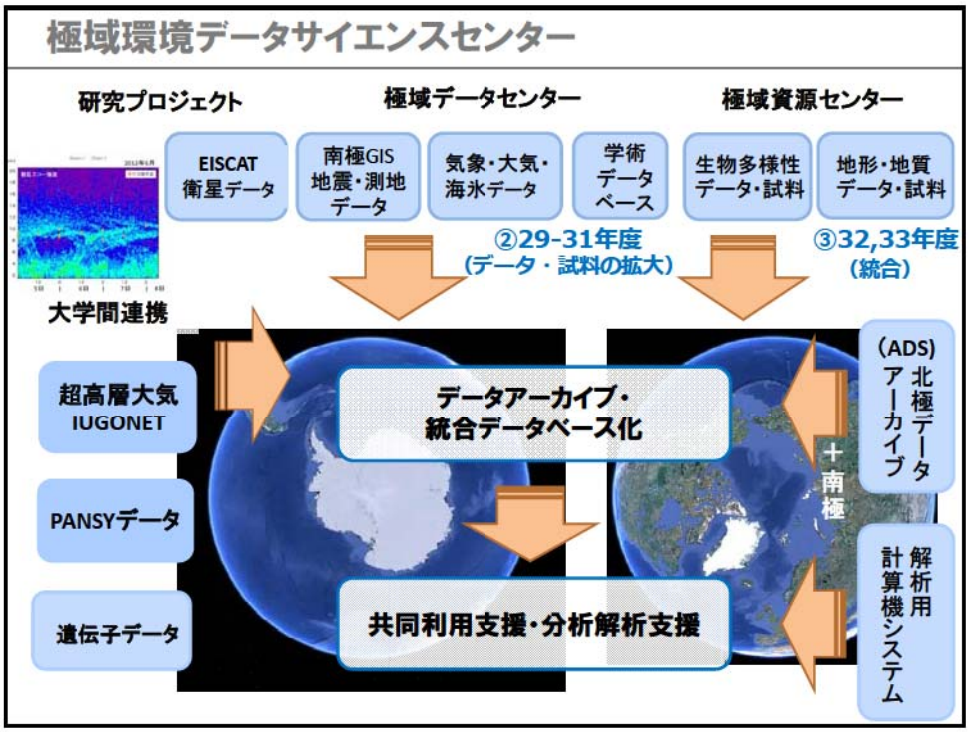
↓

セマンティクスを重視したDB統合
RDF＋オントロジー＝実用的な分散知識ベース

ドメイン内の概念とそれらの概念間の関係のセットとしての知識の形式的な表現

60


ライフサイエンス統合データベースセンター
生命科学分野のオープンサイエンス推進
データベース統合化技術の他分野への展開
科学技術振興機構との
連携プログラムと共同研究
統合化作業及び
サービスの提供
(JST/NBDC)
統合DBの
利用促進
実験データ等の
DB化・公開支援
基盤技術開発
プログラム
基盤技術
国際標準化
人材育成
連携・共同研究等
バイオインフォマティクス
DBを統合的に利用
するためのツール群
(検索と表示)
○より多くのDBを
○より高度な検索を
○より平易な言葉で
オントロジー
RDF化
用語や形式がまちまちで
多岐にわたる生命科学のDB群
一括利用できるDB群
統合化技術
の他分野へ
の
展開
国際ワークショップ
(標準化等)
DB構築者向け・利用者向け講習会
共同研究


極域環境データサイエンスセンター
研究プロジェクト
極域データセンター
極域資源センター
EISCAT
衛星データ
南極GIS
地震・測地
データ
気象・大気・
海氷データ
学術
データ
ベース
生物多様性
データ・試料
地形・地質
データ・試料
②29-31年度
(データ・試料の拡大)
③32,33年度
(統合)
大学間連携
超高層大気
IUGONET
PANSYデータ
遺伝子データ
データアーカイブ・
統合データベース化
共同利用支援・分析解析支援
+南極
(ADS)
北極データ
アーカイブ
解析用
計算機システム

社会データ構造化センター

1. 社会データ基盤整備

- ✓公的統計ミクロデータのオープン化の基盤整備（オンサイト利用，秘匿化，統計処理）
- ✓大規模社会調査による社会調査データ収集
- ✓ソーシャル・ビッグデータ共有基盤構築（自治体オープンデータ，Web予約データ，IoTセンシング）

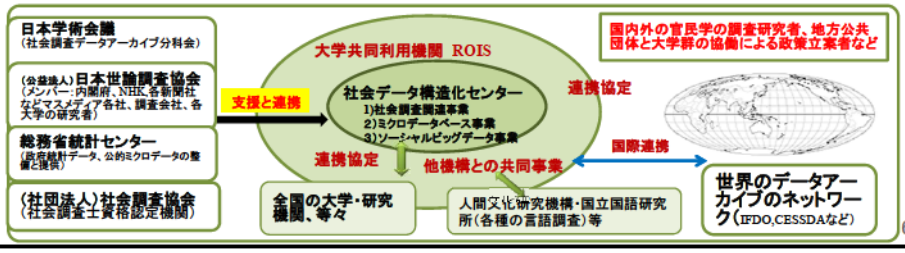
2. データ構造化の方法論構築

- ✓ データクレンジング（欠測値，異常値処理），データベースリンケージ
- ✓ 低質大規模データと高質小規模データの融合
- ✓ データ解析ツールのメニュー化事業

匿名データ提供
オーダーメード集計
オンサイト利用
国際ミクロデータベース

3. データ利用推進活動とフィードバック

- ✓ 公的統計ミクロデータ研究コンソーシアム（総務省・統計センター・大学と連携）
- ✓ 全国共同研究ネットワークの展開
- ✓ 国際公的ミクロデータWS，データ分析コンペ開催
- ✓ 社会データオープン化・共有化に関するコンプライアンス



63

低質大規模データと高質小規模データの融合

公的統計：

- バイアスなし，分散小
- 月次データ，3か月の遅延

Web 情報 (クローリング等による)

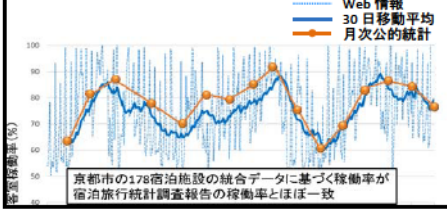
- リアルタイム情報
- バイアスあり，分散非常に大

公的統計とWeb情報の統合

- リアルタイム情報
- バイアス補正，分散減少

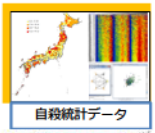
自治体等の観光政策支援

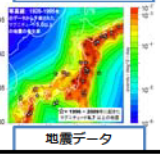
リアルタイムホテル稼働率推定（京都市）



●Webサイトに集積された「ライフログデータ」を収集し、仙台市の震災前後の宿泊施設と新幹線の復旧状況を可視化。
●平常時は、人間・社会の行動や挙動を把握して、「観光予報」などに活用し、震災時には、資源の効率的な配分を行う減災政策支援に活用。

●モバイルSNSを活用し、人間関係向上のためのソーシャルグラフとコミュニケーションログの人間・社会データ収集基盤


自殺統計データ

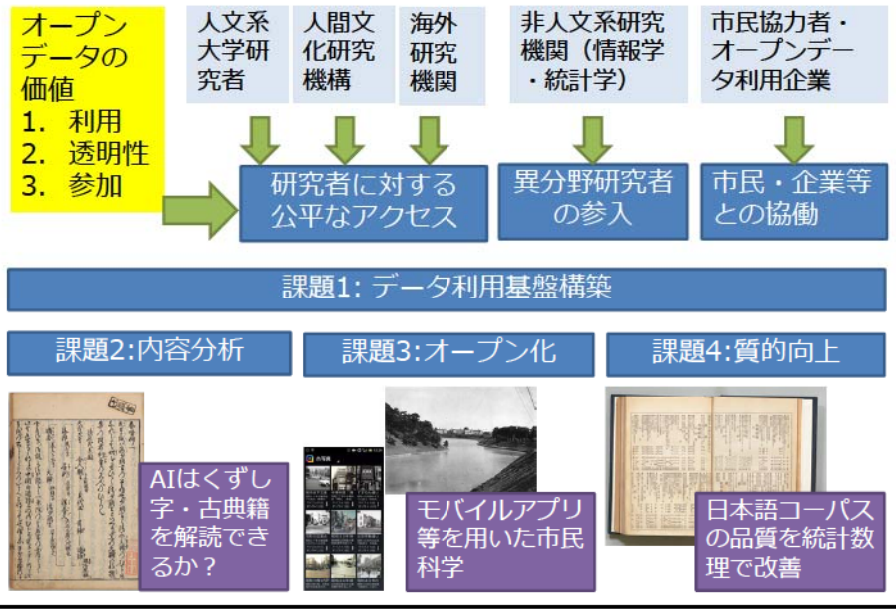

地震データ

64

人文学オープンデータ共同利用センター

- 適切なメタデータを付与して、データの検索性、再利用性を高める。
- 研究者や図書館員がオンラインで共同タグ付け（キュレーションすることで知識が蓄積するデータ評価プラットフォームを構築（国文学研究資料館）
- タグの推薦機能（メタデータや過去のタグ付け機能と知識基盤を組み合わせ、関連タグの推薦により作業効率を高める。）
- 画像の解析機能（画像データの内容分析に基づき、メタデータにないタグを推薦）
- データ評価プラットフォームの横展開
  - ✓ 人文学から自然科学まで、分野の特性を踏まえたキュレーションが行えるプラットフォームを構築
  - ✓ データのオープン化の便益の見える化

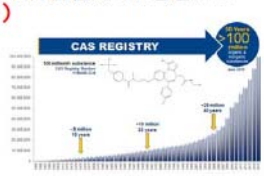
人文学オープンデータ共同利用センター



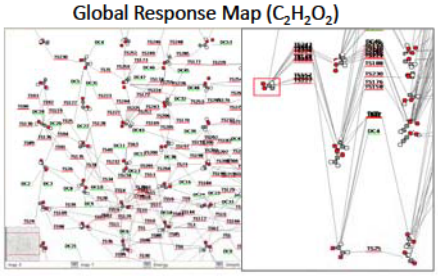
データ中心ケミストリー：　「埋蔵分子」発掘プロジェクト

既知の化合物種数 ＞ 1億種以上

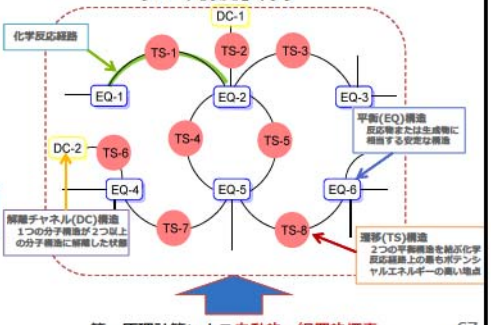
理論的に存在可能な化合物種はこれを遙かに凌駕する（**埋蔵分子**）

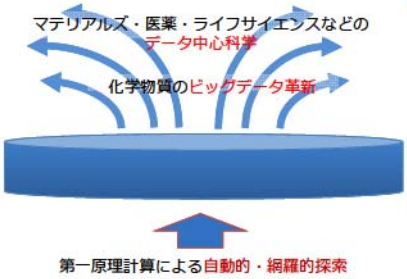


Global Response Map ($C_2H_2O_2$)



グローバル反応マップ





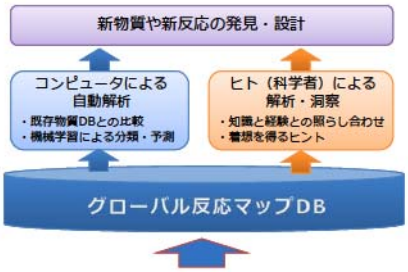
第一原理計算による自動的・網羅的探索

第一原理計算による自動的・網羅的探索

Hiroko Sato (ROIS)

67

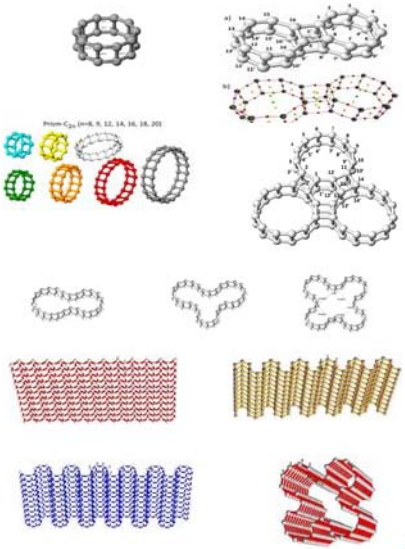
グローバル反応マップDBから「埋蔵分子」を発掘



第一原理計算による自動的・網羅的探索

新規物質と反応経路を**発見**

Discovery: Theoretically existing new materials



公開ソフトウェア
・RMapViewer ：2014年7月公開。
※2016年3月オープンソース化予定

新聞記事
・朝日新聞2015年10月19日付全国版，
「科学の扉：ビッグデータを生かす」

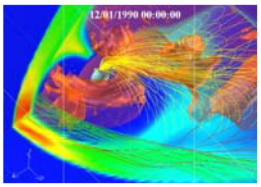
Hiroko Sato (ROIS)

68

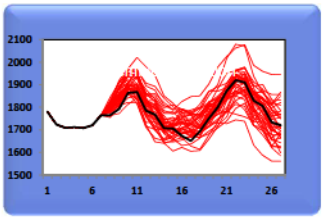
データ解析基盤 Data Analysis Platform

- データ同化
- ゲノムデータ解析
- イメージデータ解析
- 自然言語処理（メタ知識解析）
- e-サイエンス支援システム
- 可視化と構造探索
- 異種情報統合
- 機械学習
- 深層学習
- スパースモデリング

Visualization of high-dimensional data and analysis results



Spiral of knowledge development

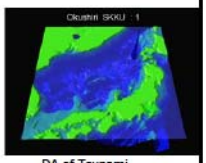


69

データ同化 Data Assimilation

統計数理研究所データ同化研究開発センター

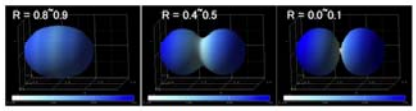
Simulation model + Observation data → Data assimilation → Realistic simulation



DA of Tsunami
中村和幸(MIMS)

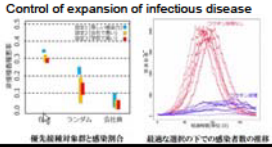
Missions:

- Development of **data assimilation techniques** such as the ensemble Kalman filter and the particle filter
- **Application** of data assimilation method to a variety of research fields such as
  - oceanography,
  - seismology, tsunami prediction,
  - magnetosphere,
  - Controlling Pandemic expansion,
  - bio informatics and biology.

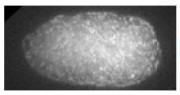


細胞質分裂時の細胞表層シミュレーション
Simulation of cytokinesis (cell division)

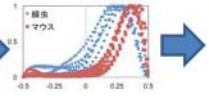
Control of expansion of infectious disease

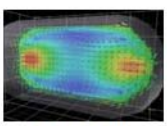


細胞質流動



せん断力分布の推定





細胞内流動の可視化

70

地震波・地震音波伝搬シミュレーション

物理モデル

- 地球中心から地表までの固体地球および地表から高度1000kmまでの大気を考慮した1次元地球構造モデル
- 気象庁が決定した震源解（モーメントテンソル解）を長さ150kmの断層に沿って配置した断層破壊モデル

各計算グリッドにおける地震波（表面波）および音波の応答波形をノーマルモード法（Kobayashi [2007]）によって計算

グリッド幅は**緯度・経度方向0.1度**および**高度方向10km**

断層を含む緯度・経度方向10度および高度方向120kmの範囲の可視化した結果を示す.

統計数理研究所が所有するスーパーコンピュータ（富士通Dシステム）を約4000ノード時間（512並列で約60時間）使用

津波よりも伝搬速度が速い音波を利用した将来型津波警報システムの構築

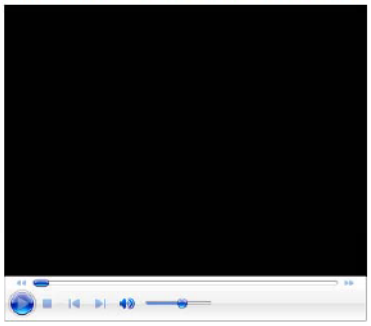
統計数理研究所 データ同化研究開発センター（樋口所長）

71

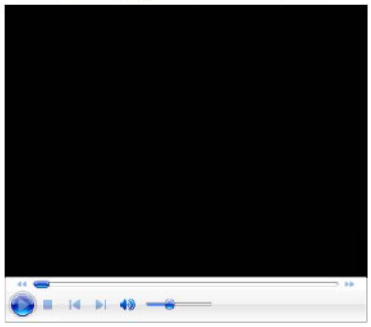
地震音波伝搬データ同化

Data Assimilation of Acoustic Propagation Cause by Earthquakes

東北地方太平洋沖地震
2011 Earthquake off the Pacific Coast of Tohoku

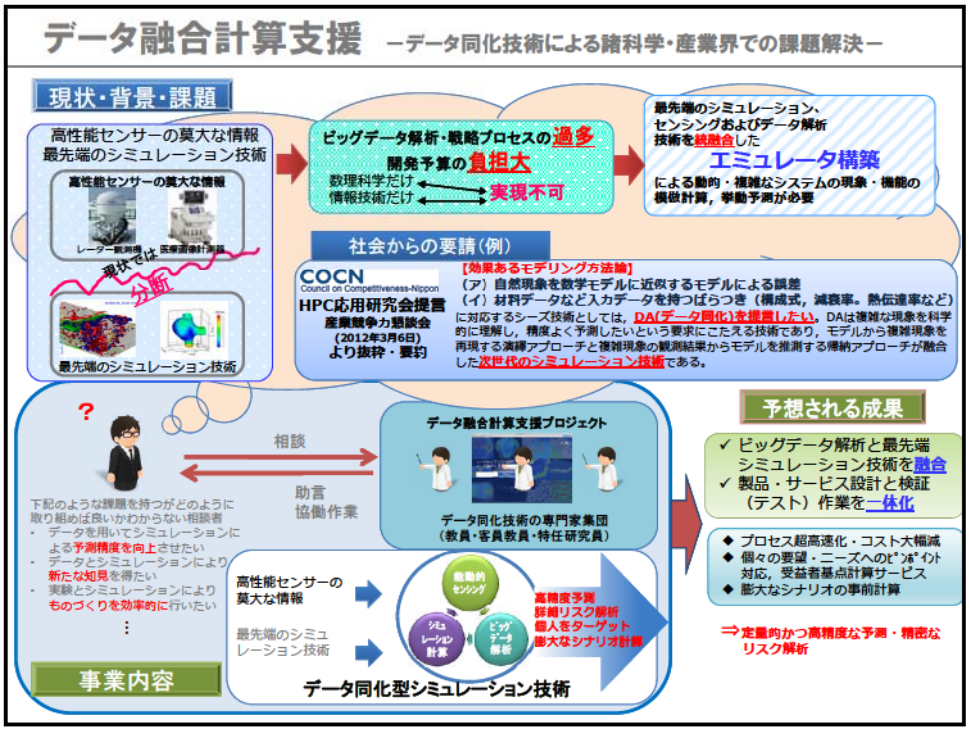


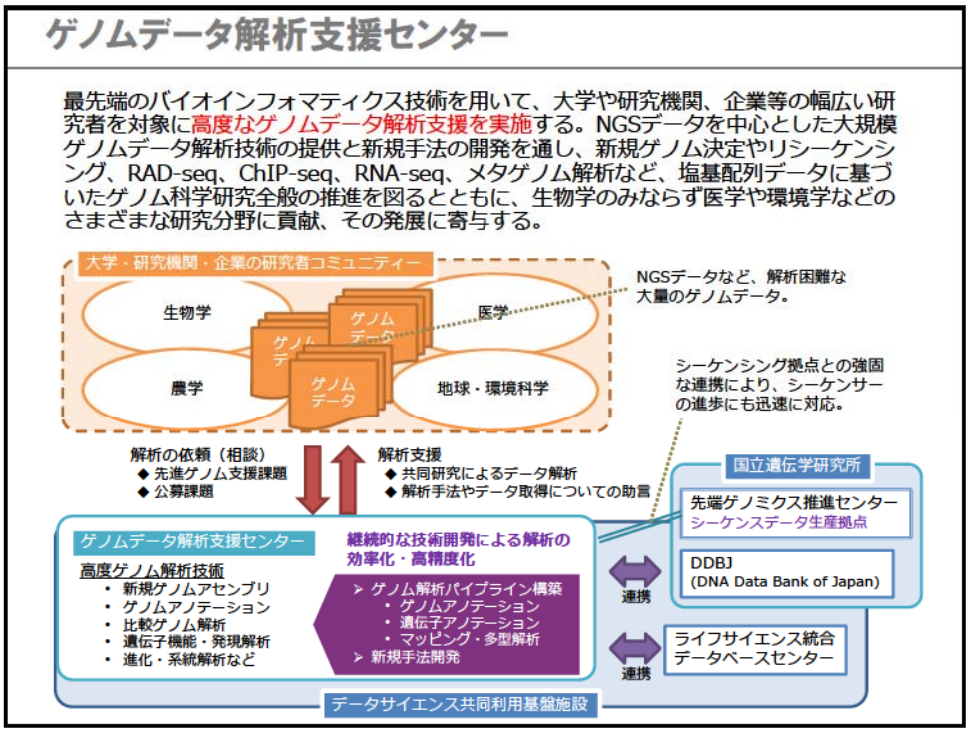
岩手・宮城内陸地震
Iwate-Miyagi Inland Earthquake



統計数理研究所 データ同化研究開発センター（樋口所長提供）

72


データ融合計算支援 ―データ同化技術による諸科学・産業界での課題解決―
現状・背景・課題
高性能センサーの莫大な情報
最先端のシミュレーション技術
ビッグデータ解析・戦略プロセスの過多
開発予算の負担大
数理科学だけ
情報技術だけ
実現不可
最先端のシミュレーション、センシングおよびデータ解析技術を統融合した
エミュレータ構築
による動的・複雑なシステムの現象・機能の模倣計算，挙動予測が必要
社会からの要請(例)
COCN
HPC応用研究会提言
産業競争力懇談会
(2012年3月6日)
より抜粋・要約
[効果あるモデリング方法論]
(ア) 自然現象を数学モデルに近似するモデルによる誤差
(イ) 材料データなど入力データを持つばらつき（構成式，減衰率。熱伝達率など）
に対応するシーズ技術としては，DA(データ同化)を提言したい。DAは複雑な現象を科学的に理解し，精度よく予測したいという要求にこたえる技術であり，モデルから複雑現象を再現する演繹アプローチと複雑現象の観測結果からモデルを推測する帰納アプローチが融合した次世代のシミュレーション技術である。
現状では
分断
相談
助言
協働作業
下記のような課題を持つがどのように取り組めば良いかわからない相談者
・ データを用いてシミュレーションによる予測精度を向上させたい
・ データとシミュレーションにより新たな知見を得たい
・ 実験とシミュレーションによりものづくりを効率的に行いたい
データ融合計算支援プロジェクト
データ同化技術の専門家集団
(教員・客員教員・特任研究員)
予想される成果
✓ ビッグデータ解析と最先端シミュレーション技術を融合
✓ 製品・サービス設計と検証（テスト）作業を一体化
◆ プロセス超高速化・コスト大幅減
◆ 個々の要望・ニーズへのﾋﾟﾝﾎﾟｲﾝﾄ対応，受益者基点計算サービス
◆ 膨大なシナリオの事前計算
⇒定量的かつ高精度な予測・精密なリスク解析
事業内容
高性能センサーの莫大な情報
最先端のシミュレーション技術
能動的センシング
シミュレーション計算
ビッグデータ解析
高精度予測
詳細リスク解析
個人をターゲット
膨大なシナリオ計算
データ同化型シミュレーション技術


ゲノムデータ解析支援センター
最先端のバイオインフォマティクス技術を用いて、大学や研究機関、企業等の幅広い研究者を対象に高度なゲノムデータ解析支援を実施する。NGSデータを中心とした大規模ゲノムデータ解析技術の提供と新規手法の開発を通し、新規ゲノム決定やリシーケンシング、RAD-seq、ChIP-seq、RNA-seq、メタゲノム解析など、塩基配列データに基づいたゲノム科学研究全般の推進を図るとともに、生物学のみならず医学や環境学などのさまざまな研究分野に貢献、その発展に寄与する。
大学・研究機関・企業の研究者コミュニティー
生物学
医学
農学
地球・環境科学
ゲノムデータ
NGSデータなど、解析困難な大量のゲノムデータ。
シーケンシング拠点との強固な連携により、シーケンサーの進歩にも迅速に対応。
解析の依頼（相談）
◆ 先進ゲノム支援課題
◆ 公募課題
解析支援
◆ 共同研究によるデータ解析
◆ 解析手法やデータ取得についての助言
国立遺伝学研究所
先端ゲノミクス推進センター
シーケンスデータ生産拠点
DDBJ
(DNA Data Bank of Japan)
連携
ライフサイエンス統合データベースセンター
連携
ゲノムデータ解析支援センター
高度ゲノム解析技術
・ 新規ゲノムアセンブリ
・ ゲノムアノテーション
・ 比較ゲノム解析
・ 遺伝子機能・発現解析
・ 進化・系統解析など
継続的な技術開発による解析の効率化・高精度化
➢ ゲノム解析パイプライン構築
・ ゲノムアノテーション
・ 遺伝子アノテーション
・ マッピング・多型解析
➢ 新規手法開発
データサイエンス共同利用基盤施設

まとめ Summary

- ビッグデータの活用が今後の学術研究及び社会の発展の鍵となる。
- 海外では，DSの教育プログラム，研究組織が急速に整備されている.
- 我が国でも提言に沿った対応が急速に始まっている.
  - データサイエンス・統計の学部，学科の新設
  - 主要６大学による人材育成事業開始
  - 統計数理研究所における棟梁レベルの育成
  - リテラシー教育～独り立ちレベルの大学教育の加速
  - 理研におけるフラグシップ・プロジェクトAIPの推進
  - JSTのCREST，さきがけ等の戦略プログラムの実施
  - コンソーシアムの形成，共通教材の開発，共通データの整備

75