データサイエンス教育と
臨床統計家育成コースへの期待

樋口知之　（情報・システム研究機構　統計数理研究所）

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

文部科学省国家課題対応型研究開発推進事業・科学技術試験研究委託事業　http://datascientist.ism.ac.jp/

ビッグデータ利活用によるイノベーション人材育成ネットワークの形成
ーデータサイエンティスト育成ネットワークの形成（平成25～27年度）

統計数理研究所 The Institute of Statistical Mathematics　東京大学 THE UNIVERSITY OF TOKYO

文部科学省委託事業 Data Scientist Training Network
データサイエンティスト育成ネットワークの形成

①ビッグデータ利活用イノベーション人材の認知度向上・**啓蒙**

イメージ◎日経就職　サクセス

1）ウェブサイト構築　2）パンフレット作成　3）シンポジウム・ワークショップの開催

②人材の**ローテーション**

東京大学情報理工系の学生を統数研で訓練後、民間企業等へインターン。人材ローテーションのパイロットから、データサイエンティストの流動性を促進するための提案等を行う。

③　**ベスト・プラクティス**の調査研究と共有

1）データ分析のベストプラクティス調査（ヒアリング等）
　●対象は民間企業・研究機関、Kaggleに登録するデータサイエンティストを予定
2）統計検定受検者にアンケート調査
3）調査レポートの公刊

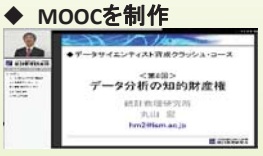
④　データサイエンティストの育成**教材の開発**

大学・企業・行政機関等で利用できるプログラムを開発
＊統数研公開講座を発展させ，データサイエンティスト向けの教材へ

⑤　**海外との連携および標準化**の検討

海外の事例を収集し，日本における取組の窓口となる
＊統計数理研究所の現在の海外機関との連携も最大限に活用
＊海外の事例調査には，IBM社ユニバーシティ・リレーションズなどの協力を得る

◆ MOOCを制作



＊インターン学生向けに、事前学習教材・インターン中の参考資料として、統計数理研究所が誇る専任教員らによるMOOCを制作
＊H27年2月には**データサイエンティスト育成クラッシュ・コースのYouTube版**を公開

◆ 報告書を発行

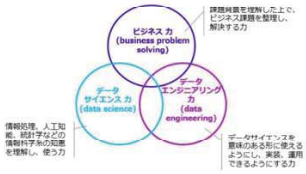
Japanese Inter-university Network for Statistical Education
統計教育大学間連携ネットワーク

一般社団法人データサイエンティスト協会

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

データサイエンティスト協会

データサイエンティストに求められるスキルセット



現場力：臨床

本コースの
育成人材像

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

T型・Π型人材の育成

横断型研究分野（Domain-General）

個別研究分野（Domain-Specific）

異分野融合の要

統計数理：事例1

情報学：事例2

横幹連関係学会
電子情報通信，情報処理，制御，人工知能，シミュレーション，OR，応用数理，品質管理

音声　人工衛星　バイオ　地球科学　自然言語　経済　未適用・未利用分野

近未来の問題の解決

細分化された個別科学をつなぎ融合

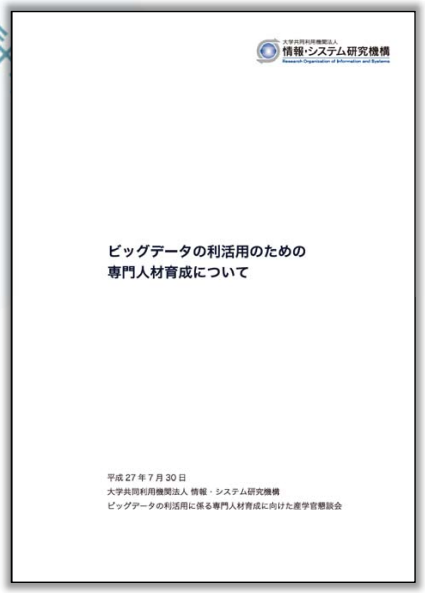
T型人材，Π型人材の育成

データ中心科学
個別科学

データ中心科学
個別科学　個別科学

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 The Institute of Statistical Mathematics

ビッグデータの利活用のための専門人材育成について



平成27年7月30日
大学共同利用機関法人 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

大学共同利用機関法人 情報・システム研究機構 Research Organization of Information and Systems
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

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
出 席・陪 席 者 一 覧

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
陪席	栗辻 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


座長　北川源四郎


委員　安宅和人


委員　岡本青史
（その後　人工知能研究センター長）

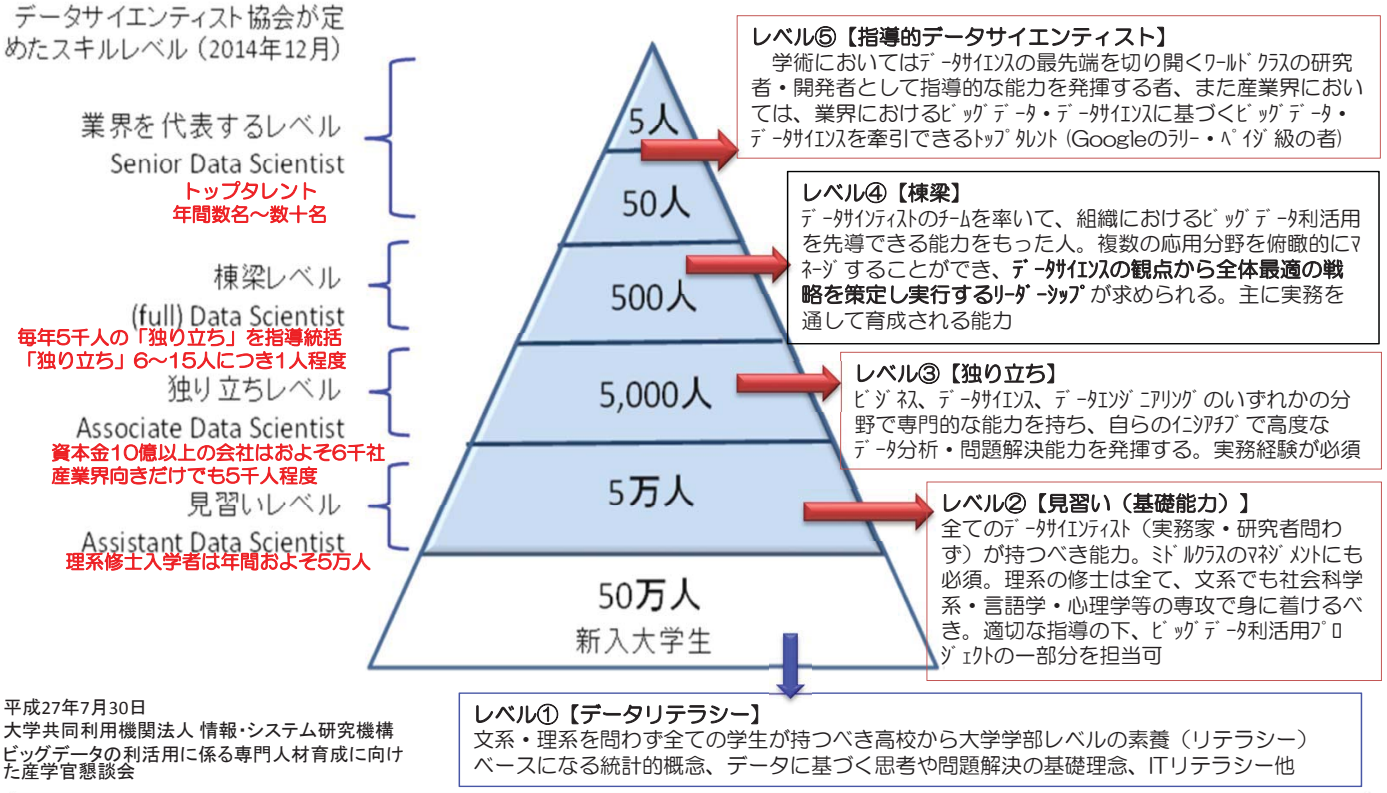

委員　佐藤 俊哉


委員　渡辺美智子

大学共同利用機関法人 情報・システム研究機構 Research Organization of Information and Systems
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

データサイエンス人材育成のあるべき姿と現実に向けた仮説

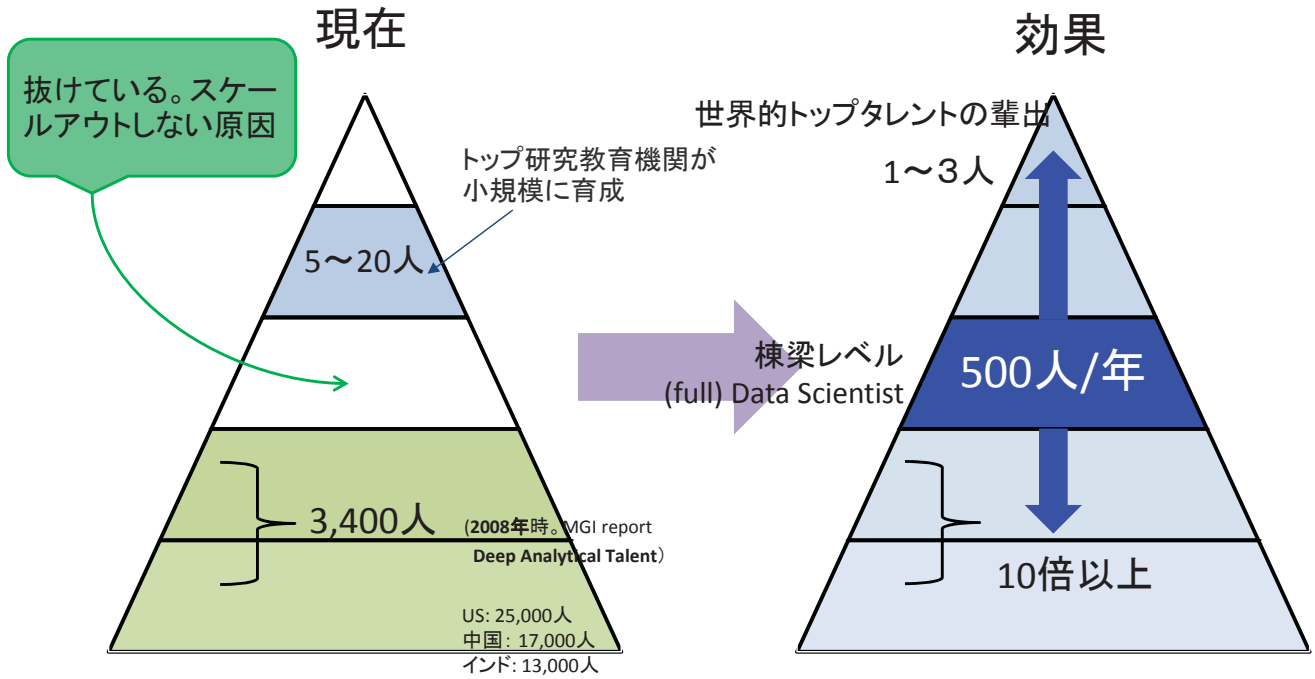
—育成レベルと、毎年の育成目標人数—



平成27年7月30日
大学共同利用機関法人 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

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

データサイエンティスト育成のあるべき姿と実現に向けた仮説の提示



育成が遅れ、最も欠けている棟梁レベル(左)と育成実現時の波及効果(右)

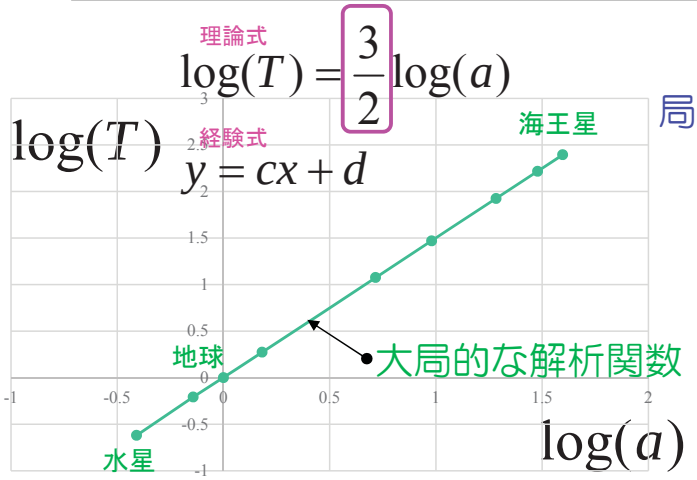
大学共同利用機関法人 情報・システム研究機構 統計数理研究所

概括：帰納法とデータ今昔

ケプラーの第三法則

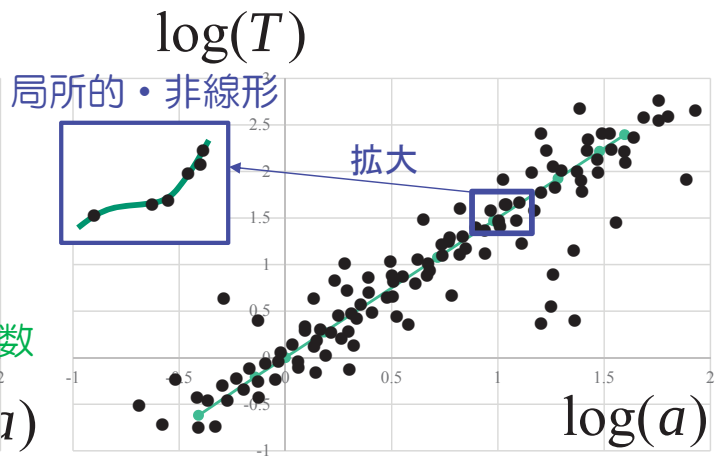
師匠であったティコ・ブラーエの観測記録から推定し定式化

惑星の公転周期 T の2乗は、楕円軌道の半長軸 a の3乗に比例する



ティコ・ブラーエの観測データ

狙って観測、帰納推論から経験則、そして一般則（万有引力の法則）を導出



ビッグデータ

無目的・副産物的にデータが蓄積され、経験則のみでOK（目的ー**予測・判別**ーが達成）

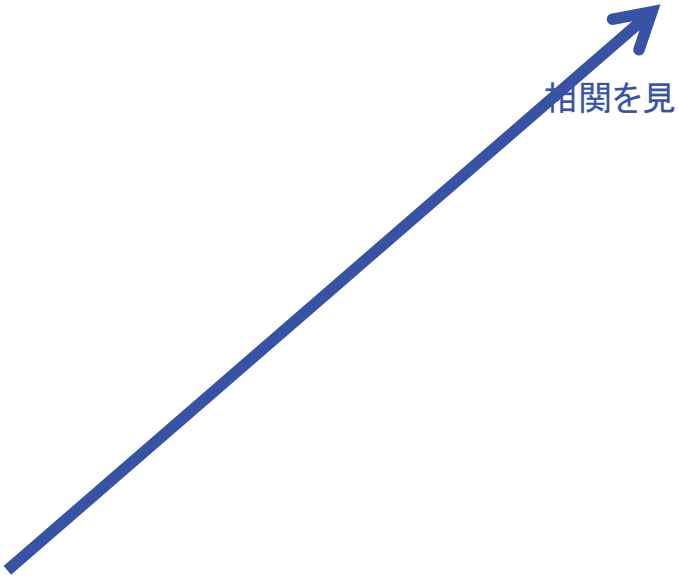
ビッグデータ時代以前：**良質な空間を見つけること**
1）説明変数の選択　2）線形性　3）少ないパラメータ

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

帰納法の弱点①：相関と因果

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

ノーベル賞輩出率とチョコレート消費量



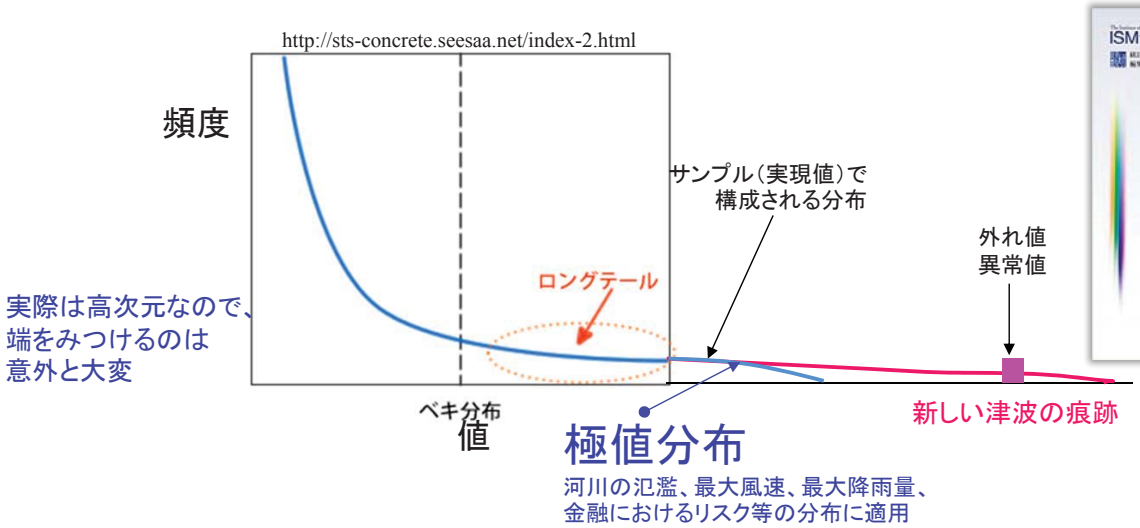
1000万人あたりのノーベル賞受賞者数

チョコレート消費量

The New England Journal o f Medicine (October 10, 2012)
Chocolate Consumption, Cognitive Function and Nobel Laureates
By Franz H. Messerli, M.D.

帰納法の弱点②：不（未）観測とデータの質

サンプリングの問題



極値分布

河川の氾濫、最大風速、最大降雨量、
金融におけるリスク等の分布に適用

『端にこそイノベーションの卵』

- 新発見、ひらめき
- クレーム（PL法対応） ←エラー、故障、不正、侵入

✓ データの偏り
✓ そもそも観測できていない？

研究不正の増加

発見・新しい知識

その信頼性と頑健性を担保

これまでは・・・
学術雑誌での論文審査の仕組みで確立

人類の英知

学術分野超細分化による弊害の発生！
⇒先端的成果の理解の困難
・・・高度な専門的知識/長年の経験が必須

学術雑誌での論文審査の仕組みでは対応しきれない！！

① グローバルな研究開発競争の過熱
② 競争を公正化するための処方としての論文引用数のみによる評価システムの一般化

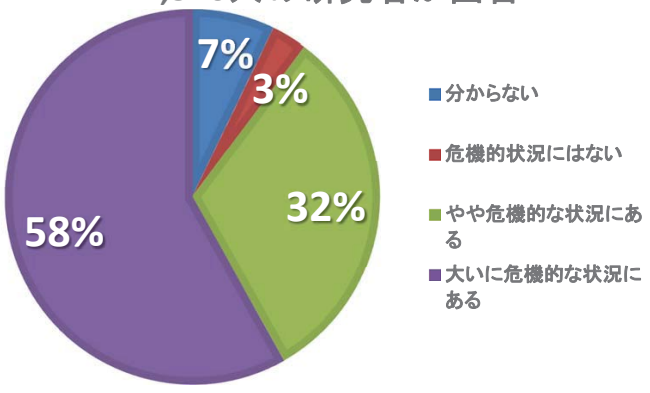
研究不正発覚の増加

学術成果の検証性に関する構造的な脆弱店を組織的かつ巧みに突く

科学論文の再現性の危機感

Nature誌2013年8月 実験結果を再現できない重要な研究論文がコンスタントに大量に発表されている分析記事の発表

「医学生物学論文の70～90%が再現できない！」



©「再現性の危機」はあるか？−調査結果− Nature ダイジェスト Vol. 13 No. 8 | doi : 10.1038/ndigest.2016.160822
原文：Nature (2016-05-26) | doi: 10.1038/533452a | 1,500 scientists lift the lid on reproducibility
http://www.natureasia.com//ja-jp/ndigest/v13/n8/「再現性の危機」はあるか？%26minus%3B調査結果%26minus%3B/77048

研究の再現性：学術界の対応

【Nature誌】
＊2013年5月から編集方針を変更
- Method欄の文字数制限を撤廃
  実験と解析に関しては十二分に説明してもらうため
- データ解析に関する統計家の意見を重要視
  解析や解釈の恣意性を排除するため

【Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results】
＊ 2015年2月開催米国国立科学財団(NSF：National Science Foundation)の支援によるワークショップ
- 統計的観点からの再現性の問題がテーマ

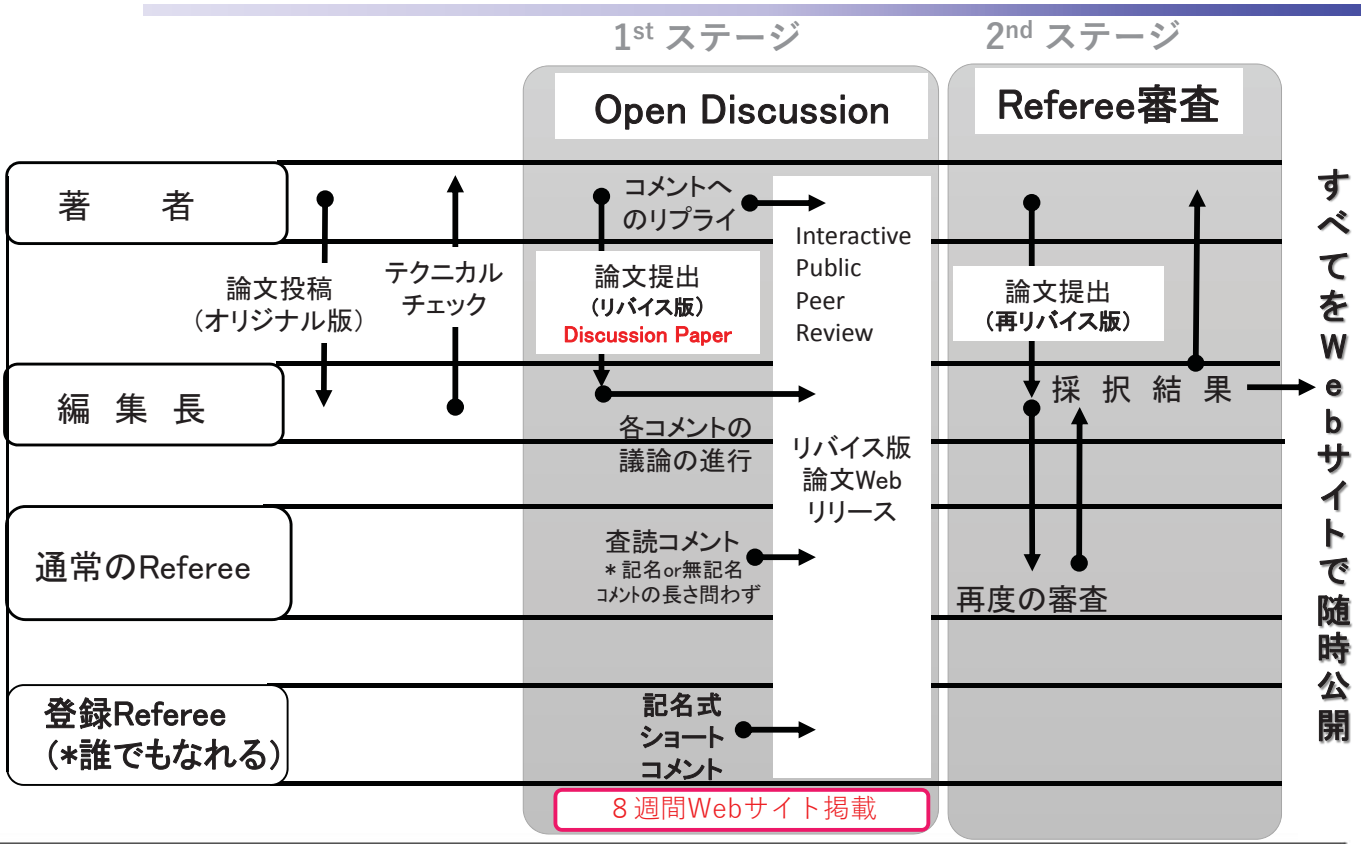
【JASA(Journal of the American Statistical Association)】
＊2016年9月から編集方針を変更
- Associate Editor for Reproducibility (AER)を編集委員会への追加

通常の審査を経て出版が決まっている論文に対し科学的価値の観点からさらに審査を行う

他人による再解析（再計算）の実効性にまで踏み込む先駆的な取組

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

Nonlinear Process in Geophysics(NPG)の査読システム



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

経済統計データの信頼性と官庁データサイエンティスト

4.「官庁データサイエンティスト」の育成・各府省の支援（総務省統計研修所）

① 研修プログラムの充実強化を図り、**「官庁データサイエンティスト」育成**を推進し、各府省における**経済統計改善技術の向上**を図るとともに、**ＥＢＰＭ**（Evidence Based Policy Making）**の環境を整備**

② ビッグデータの利用等、**高度な統計技術の研究開発**、各府省への支援の強化

統計家の行動基準

【日本計量生物学会「統計家の行動基準」2013年制定より】

「統計家は，データを収集し，統計手法を用いて不確実性の程度を定量的に明らかにしたうえで結論を導き，科学，医学，経済，社会などのさまざまな領域における意思決定に関与し，人々の健康や安全，福利の増進や環境の保全，社会の経済の安定と発展に貢献する専門家である」

【国際統

「統計家の行動基準に則り、適正な研究が実施できる臨床統計家の育成のためのコース」

			(RSS : Society)
1949年		倫理規定策定開始 Ethical Guidelines for Statistical Practice	
1979年	倫理規定策定開始　Professional Ethics		
1985年	**倫理規定制定　Professional Ethics**		
1989年		**倫理規定制定** **Ethical Guidelines for Statistical Practice**	
1993年			**規程制定　Code of Conduct**
最新版	**2010年7月** https://www.isi-web.org/index.php/activities/professional-ethics In carrying out his/her responsibilities, each statistician must be sensitive to the need to ensure that his/her actions are, first, consistent with the best interests of each group and, second, do not favor any group at the expense of any other, or conflict with any of the Principles.	**2016年4月** https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf Good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations.	**2014年6月** http://www.rss.org.uk/Images/PDF/join-us/RSS-Code-of-Conduct-2014.pdf This code of conduct has been drawn up to reflect the standards of conduct and work expected of all practicing statisticians.

個人への直接的な関与：Personalization

Achieve technologies that personalize products and services

personal, unique, individual, characteristic

個人、個性、個別、固有

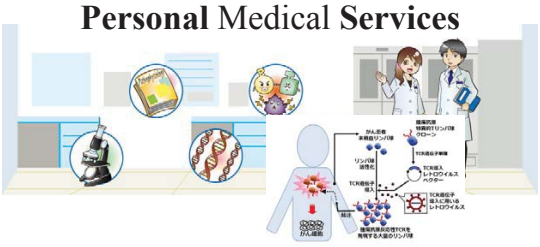
条件付き分布の構成 Conditioning

$$p(\mathbf{y} \mid \theta_1, \ldots, \theta_P)$$

Personal Services

Real-Time Bidding

Personal Medical **Services**



技術のハードル vs. 顧客のハードル

（森川教授＠東大 談）

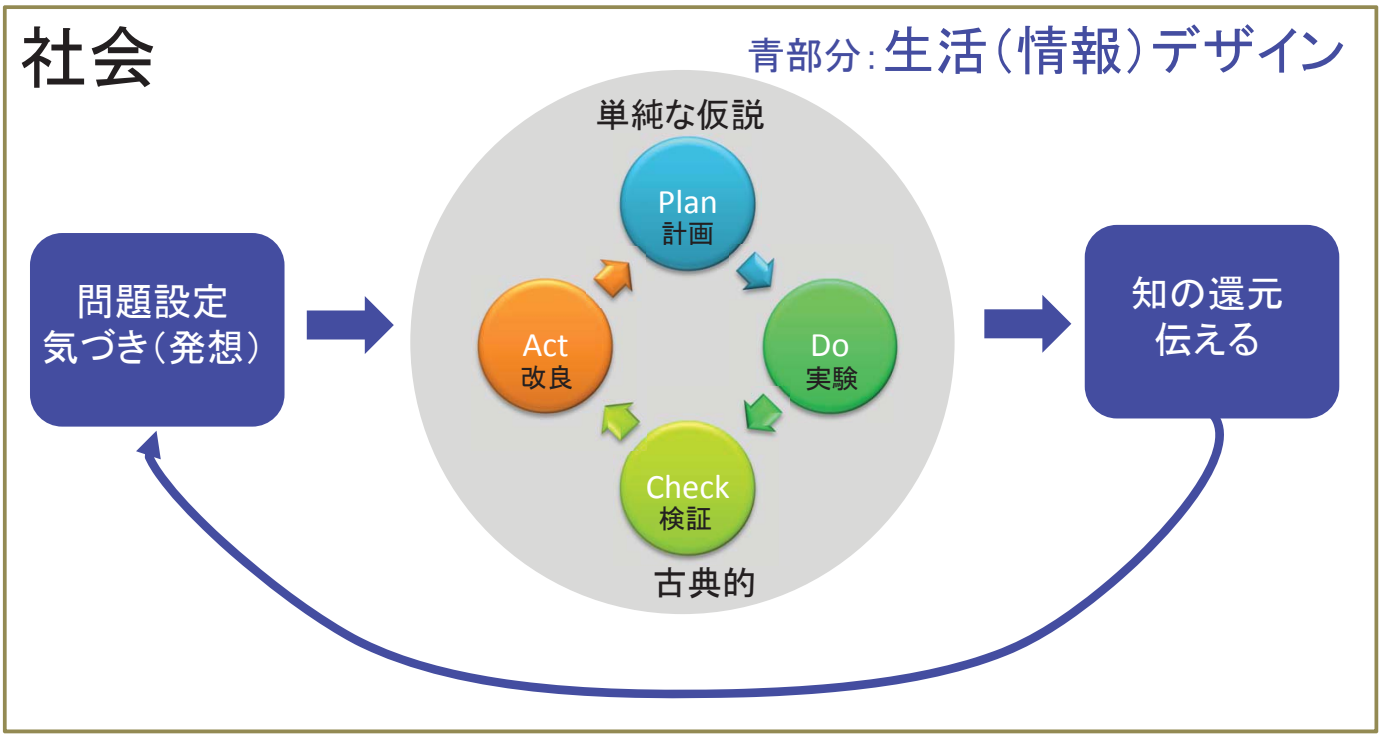
30年前

今

グーグル・グラス、モバイク（中国）

PDCAサイクルだけでは不足・時代遅れ

PDCAサイクルに加えてこれからわれわれはどうすべきか？



参考文献

樋口知之, データ駆動科学技術を担う人材の育成：確率的思考と逆推論, *情報管理*, Vol.59, No.1, 53-56, 2016.
樋口知之, 人工知能はみようみまねマシンの究極形, *情報管理*, Vol.59, No.5, 331-335, 2016.
樋口知之, シチズンサイエンス　相互監視それとも共進化？, *情報管理*, Vol.59, No.9, 629-635, 2016.

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