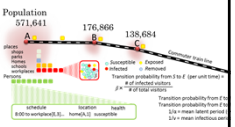
総	会		選	出
第３６回通常総会 より				

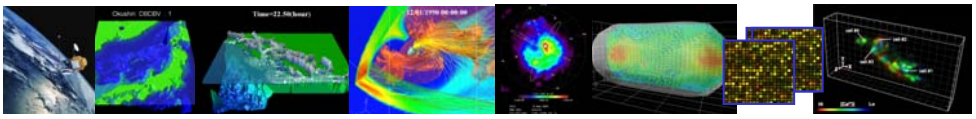
人への投資を触媒とする ビッグデータからの価値創出

樋口 知之
(情報･システム研究機構　統計数理研究所)

人への投資を触媒とする
ビッグデータからの価値創出

樋口知之　（情報・システム研究機構　統計数理研究所）





アウトライン

1. ビッグデータとは
2. アナリティクスの4つの落とし穴
3. 予測とモデル
4. 4つの利用レベルのステージ
5. 日本固有の問題点
6. 対応策：　人材育成

1．ビッグデータとは

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

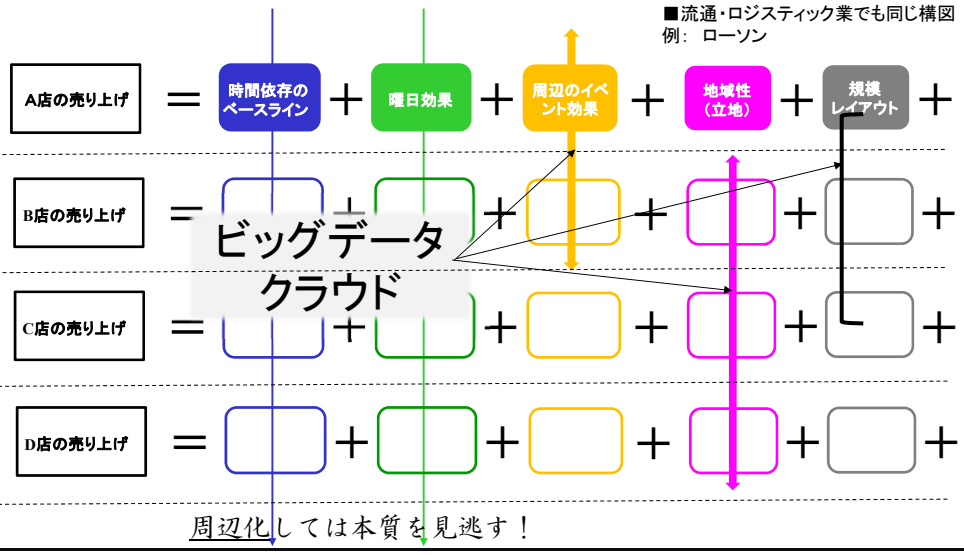
3V: 量(Volume)、種類(Variety)、頻度(Velocity)

5V: 価値(Value)、情報の正確さ（信憑性）(Veracity)

要因分解モデル：バラバラを癖で束ねる

小売りの売り上げ予測

■流通・ロジスティック業でも同じ構図
例： ローソン



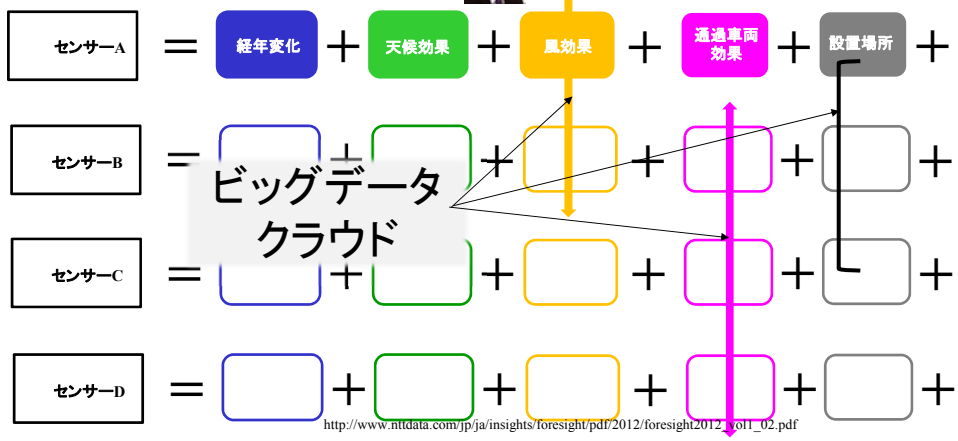
周辺化しては本質を見逃す！

統計数理研究所

要因分解モデル：システムの「人間ドック」

橋梁センサー等の社会インフラ維持・管理用モニター

NTTデータ、NEC（次世代航空機着陸システム開発）
GE（Industrial Internet）



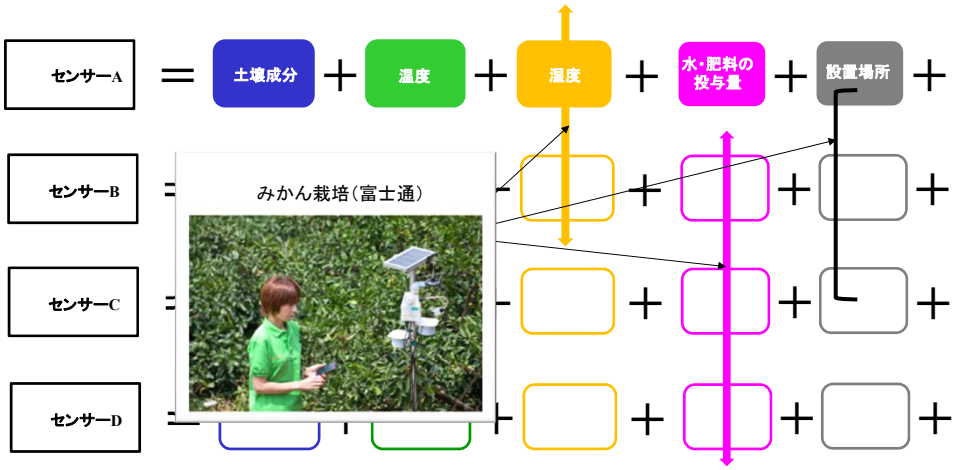
http://www.nttdata.com/jp/ja/insights/foresight/pdf/2012/foresight2012_vol1_02.pdf

統計数理研究所

要因分解モデル：匠を超える技をつくる

第一次産業（農業、漁業）を変える、再生可能エネルギー（風力発電）を変える

オープンデータ化の動き（地力保全土壌図データ、…



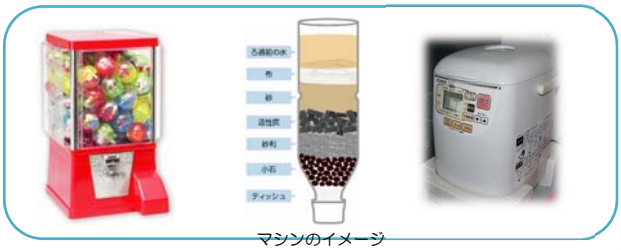
『マシンに入れれば何かでてくる』は幻想

■よくあるケース

「ビッグデータはいろいろ社内にあるのだけれども、先生、何かできませんかねぇ？もったいないといつも思っているのですが。」

「この種の発言をされるお客さんの案件はお断りするようにしている。営業にはいやな顔をされるが」

某ビッグデータ･コンサルティング担当マネージャー



マシンのイメージ

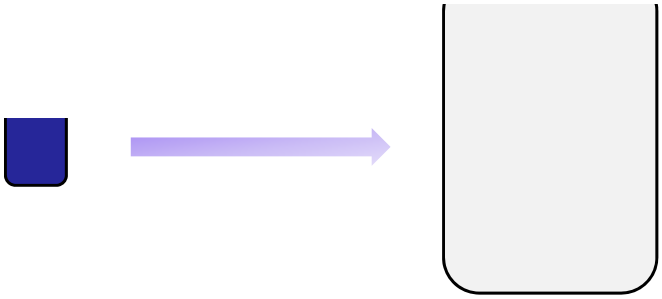
ビッグデータ環境下における研究開発推進の鍵

ビッグデータは「価値密度」がかなり低い

（統計数理研究所　丸山；　PFI 岡野原　談）

価値（←目的による！）密度の定義

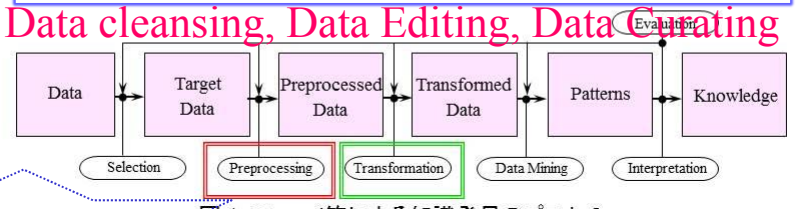
$$\text{価値密度（目的）} = \frac{\text{価値総量（目的）}}{\text{データ総量 Volume}}$$



多くのプロセスに分解

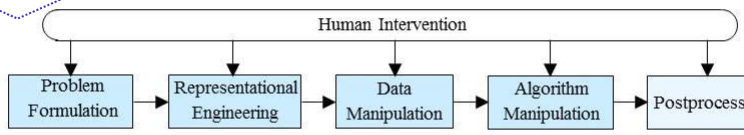
抽象物を対象にすると
誰もプロセスに分解できない！

Data cleansing, Data Editing, Data Curating



13年前の資料から

図 1: Fayyad等による知識発見のプロセス



データ処理の流れ 図 2: Langleyによる知識発見のプロセス

発見科学研究 成果報告集
宮野悟(平成13年3月)

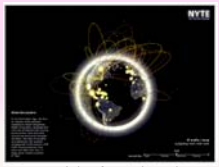
Big data techniques and technologies

- **TECHNIQUES FOR ANALYZING BIG DATA**

統計科学、数理工学

- **VISUALIZATION**

Tag Cloud



Spatial Information Flow

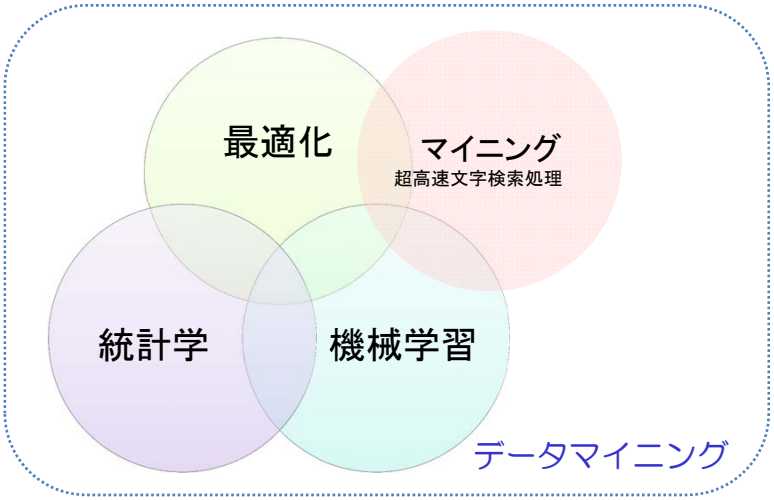
McKinsey Global Institute
June 2011
Copyright © McKinsey & Company
www.mckinsey.com/mgi

- **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

３ or ４大　解析要素技術分野



２．アナリティクスの『４つの落とし穴』

ビッグデータの操作に没頭して本質を見逃す危険性

スモールデータの取り扱いの十分な理解
無くしてビッグデータによる成功ありえず。

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

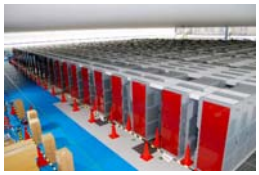
$$\max . f(\theta) \quad \theta' = (\theta_1, \ldots, \theta_p)$$

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。

異質性
セレクションバイアス：
販促のマイナス効果（価格プロモーションとリピート購買率）、ウェブサイトのページビュー数、顧客ロイヤリティとCRM
（阿部、2013）

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。**

率（頻度，指標）がくせもの

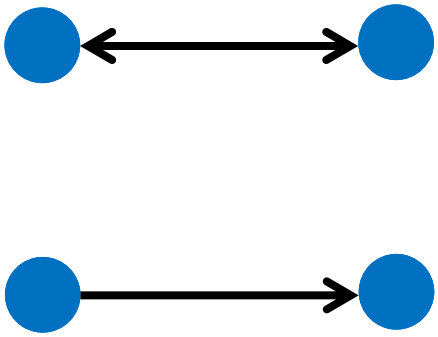
$$\text{リピート購買率} = \frac{a + b}{A + B}$$

コンバージョン率
クリック率

A: ターゲットの総数
a: アクション（購買）数
B: 興味のない集団の人数
b: アクション（購買）数

既存顧客 (A, a) vs. 新規 (B, b)

落とし穴2：列挙処理、相関と因果



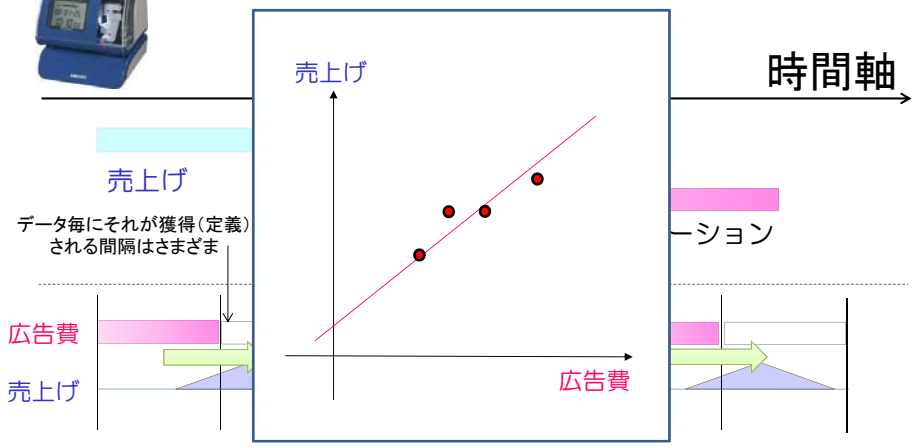
言っている内容は正しいが，
言い方は誤解を与える。

ビッグデータの時代には、暮らし方から世界との付き合い方まで問われることになる。特に顕著なのは、相関関係が単純になる結果、社会が因果関係を求めなくなる点だ。
「結論」さえわかれば、「理由」はいらないのである。

Granger causality

2003 Nobel Memorial Prize in Economic Sciences (from Wikipedia)

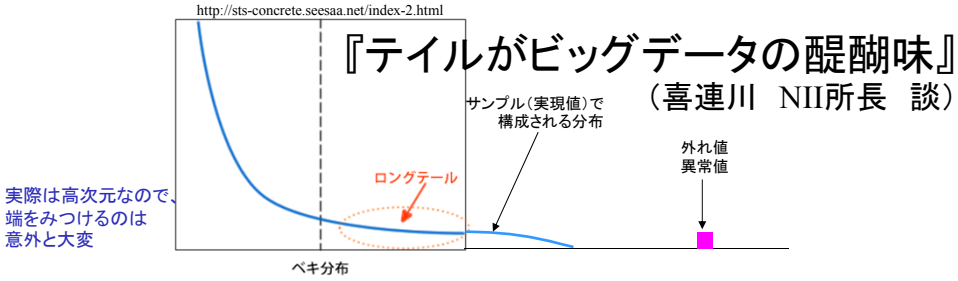
- データのタイムスタンプとラグの最適化
- 介在効果のモデル化



統計数理研究所

落とし穴3：全てのデータを取り扱う意味

帰納法の弱点



『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）

そうでなければサンプリング（標本抽出）によって一部のデータを分析することで十分（費用対効果を最初から考えること）

統計数理研究所

ボナンザ（保木 氏作）

2006年5月　第16回世界コンピュータ将棋選手権大会優勝

論理思考とデータ解析の組み合わせ

探索：巨大な状態空間の中の効率的、効果的なカづく探索

○**全幅検索**と**選択検索**のハイブリッド

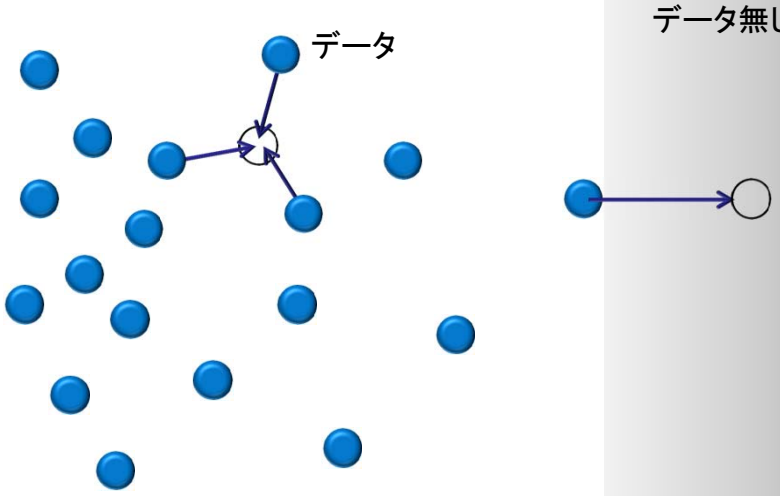
機械学習：6万局の棋譜データ（※）から、評価関数のパラメータを自動生成。↑の“効果的”の源泉。

※プロの公式戦の対局3

歩		歩		
	歩	銀	歩	歩
		玉	金	金
香	桂			

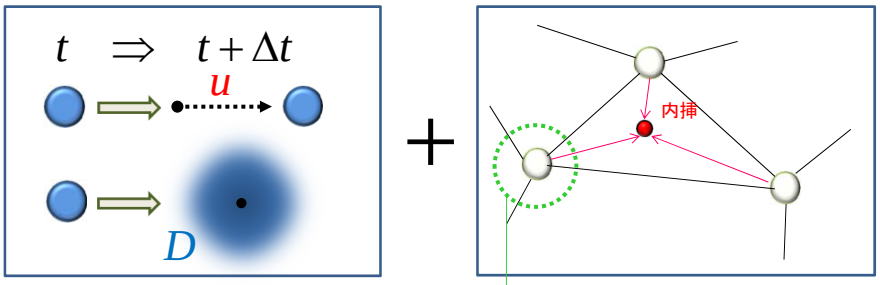
ボナンザ囲い

落とし穴4：内挿と外挿問題



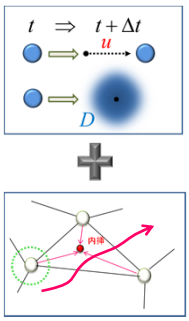
いろいろなビジネス展開が可能

移流と拡散 ＋ **クラウドソーシング** ＝ 予測能力
フォワード計算モデル　　現況を捉える認識力 スマホ(とGPS)



情報の不確実性
（多人数からのレポート：多様なノイズの混在）

モバイル空間統計（NTTドコモ）



http://www.nttdocomo.co.jp/corporate/disclosure/mobile_spatial_statistics/

3．予測とモデル

予測能力の向上に集中！

二つの異なる推論法

二つの機能の性能の合成

フォワード（前向き）計算モデルの記述力
＋
対象の現状態（現況）を捉える認識力
＝　予測能力

計測手法のイノベーションと直結

予測モデルがないと…

「計算モデル無き予測法は予測とはいえず。」
渋滞情報を流すと、人の多くは同じ行動をとりさらに状態が悪化

“手なり予測”（リクルート・テクノロジー　西郷氏　談）
手なりとは：　手役などを考慮に入れず、牌効率のみでアガリに向かうこと。

物流：　急な対応　（人の動員）
小売り：　明日の量的な販売予測。商品毎の売り上げ予測。
　　廃棄をいかに少なくするか？
　　廃棄率（廃棄リスク）
飲食：　アルバイト（非常勤）の日々、状況毎の必要人数
製造：　故障の前にお知らせ（車、ガス機器、コピー機）

Nowcasting と Forecasting
プローブデータの活用：通常時にはきわめて有効

出展：国土技術政策総合研究所　上坂克巳氏のスライドより
http://www.nilim.go.jp/lab/bbg/saigai/h23tohoku/houkoku2/happyou/09.pdf

Nowcastingからダイナミックモデルを作成する

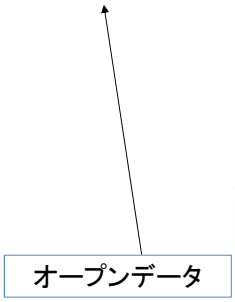
現在のCPSはNowcasting

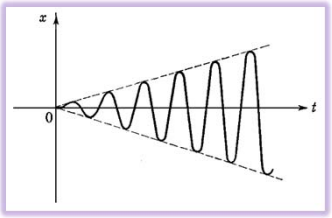
NHK「震災ビッグデータ」

東大生産研・関本准教授　スライドより

プローブデータとITSの方向だけで大丈夫か？

予測なき情報提供は渋滞を悪化させることも多い





http://www2.toyota.co.jp/jp/news/13/05/nt13_0511.html

位相差（時間差）のある系は、
ゆらぎの振幅を増大させる。
人の意志決定は画一的ではないが、同じことを考える割合が多いと問題が生じる。

先読みナビ機能が必須

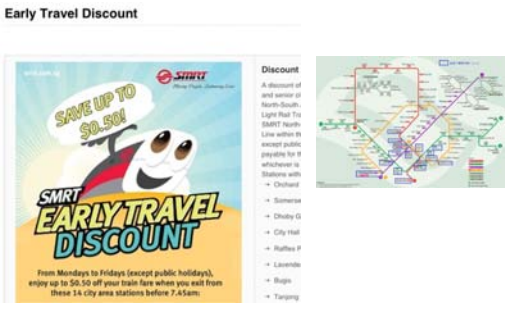
公共サービスとビッグデータ

■通常時： Nowcastingが中心
- オープンデータ＋クラウドソーシング
  自然とデータが集まる仕組みづくり
- 民間との協働
  ビジネスモデルの創発はやらない。「武士の商法」

■非常時： Forecasting機能が大切
- ITインフラのダメージ、バックアップシナリオの事前想定
- 先読み情報サービスがはずれた場合のリスク：訴訟、一般からの批判
- トリアージが肝

シンガポールの交通システム

Electronic Road Pricing
（ETCのようなもの。1SD=78円）

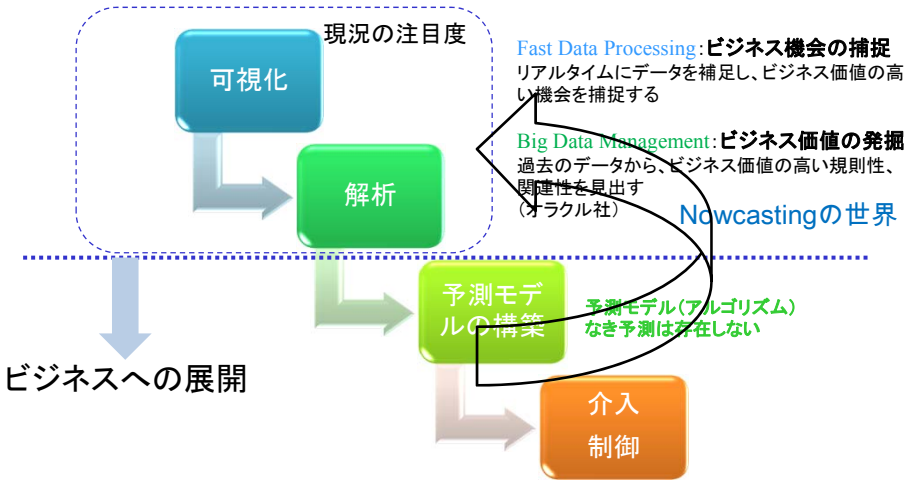


- 平日始発～7：45まで中心部で降りる乗客の運賃を試験的に無料化。これによりピーク時の利用者は平均**7%**減少。
- 地下鉄運営会社はピーク時の混雑を**10～20%の緩和**を目指すとして、早朝無料を今後1年間続けるとのこと。
- ラッフルズ・プレイスやシティホールなど中心部の16駅で、7時45分から8時までの間に改札を出た乗客も0．5シンガポール（S）ドル（**約40円**）を割り引く措置も合わせて行う。

価値？
ヒューマンモデルの欠如

４．４つの利用レベルのステージ

ビッグデータ利活用の４つのステージ



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

「広告」から「個告」の時代へ
（日経ビジネス 2013.09.30号）

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！

帰納と演繹

理論と仮定から結果を導く

vs.

結果から原因を探る

帰納と演繹

理論と仮定から結果を導く

vs.

結果から原因を探る

あらゆる場面がオークション（RTB）化

個人のライフスタイル・嗜好を読み解く

- デジタル値札
- 赤外線センサー

映画：『マイノリティ・リポート』の1シーン

2002年のアメリカ映画SF映画。ドリームワークス作品。
ある日、犯罪予防局の刑事ジョン・アンダートン(トム・クルーズ)は、ジョンによる殺害が予知されたことで追われる立場となってしまう。

網膜認識を利用してユーザー特性に合わせた情報を発信する.
（“個” 人認識型　デジタルサイネージ）

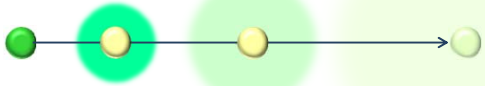
POSデータ ＋ センサーデータ →
個人毎の潜在価格を読み取ることは可能

- バンドル売り
- 電子チラシ
- クーポン

Sparse, Sporadic, and Spontaneous Society

（疎）　（散在的）　（自発的な）

社会構造の大きな変化
“もの”



システム

薬、服、靴
（3次元スキャナー、3次元プリンター）

サービス
- 双方向性
- 受益者起点
- 制御可能性

双方向性機能が弱いビジネスモデルの価値

受益者起点を支えるのは双方性機能。それもマシンで実現。

メディア

広告

もっとも双方向性が高いのはソーシャルゲーム

B2B2….2B2C

B2C

下流からスタートせよ：ただし逆流は難しい

受益者（消費者）を意識する
ことからはじめよ！

「モノ」の価値

- 商品よりもデータの囲い込み（リンケージ）が重要
- すべてのモノを情報収集端末と捉え、サービスを含めてどんなビジネスモデルを描くかが肝 → 車→プローブデータ

出展：国土技術政策総合研究所　上坂克巳氏のスライドより
http://www.nilim.go.jp/lab/bbg/saigai/h23tohoku/houkoku2/happyou/09.pdf

- モデルの読み合いに負けたら終わり（オンライントレーディングを考えてみると分かる）

ロンドンでの概念実証プロジェクト

Presence Orb 「現実世界のクッキー」

「われわれは匿名化され集計されたMACのデータを収集している。つまり、個人や個々のMACアドレスを追跡しているわけではない。ORBは、容器周辺のすべての足取りを3分間収集して、匿名化され集計された1件の報告を各サイトから送信している」

人もプローブ！

（2013年8月13日　MSN産経ニュース）

5. 日本固有の問題点

企業上層部のデータリテラシー

○人材育成（人材争奪戦）
○法体系整備、プライバシー

- 個人情報保護や企業技術の秘匿の観点から，企業内データを利用した外との共同研究を一顧にしたがらない上層部の態度も問題。
- 企業内のデータサイエンティストに社内データの持つ価値を検討させ，ケースバイケースで外の力を借りて新しい製品・サービスの開発のチャンスが多くうまれることを深く認識すべき。

ビッグデータを開示する正の面に
もっと目を向ける。
ただし，ケースバイケース

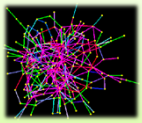
個人情報保護とデータの本格的利活用

One for All, All for One

パーソナルデータに関する検討会
＜事務局＞
総務省　情報流通行政局 情報流通振興課
経済産業省　商務情報政策局 情報経済課
※　内閣官房　情報通信技術（IT）総合戦略室

データとビジネスの関係

日本固有の弱さ



「もの」から
「システム」へ

匠とマシン

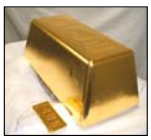
無形物にお金を
払わない文化

説明済み

ビッグデータは巨大なゴミ箱？

生ゴミ　プラスチック　新聞・紙　ビン、アルミ缶

分別、整理することで



ビッグデータの実際は、**そのまま**だと単なる屑の山

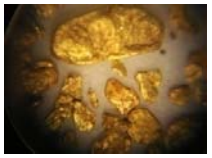
異常値の混在，欠損の頻出，フォーマットの変更，測定環境の変化等々

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



マネーボールの実在

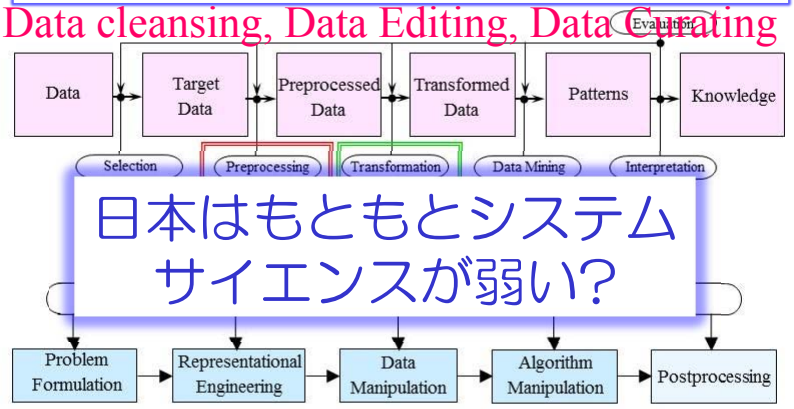
「もの」は分解できる

http://www.ifixit.com/Teardown/iPad-Teardown/2183/1

成功の鍵：データ・サイエンティストの抱え込み

抽象物を対象にすると
誰もプロセスに分解できない！

Data cleansing, Data Editing, Data Curating



日本はもともとシステム
サイエンスが弱い?

データ処理の流れ 図 2: **Langley**による知識発見のプロセス

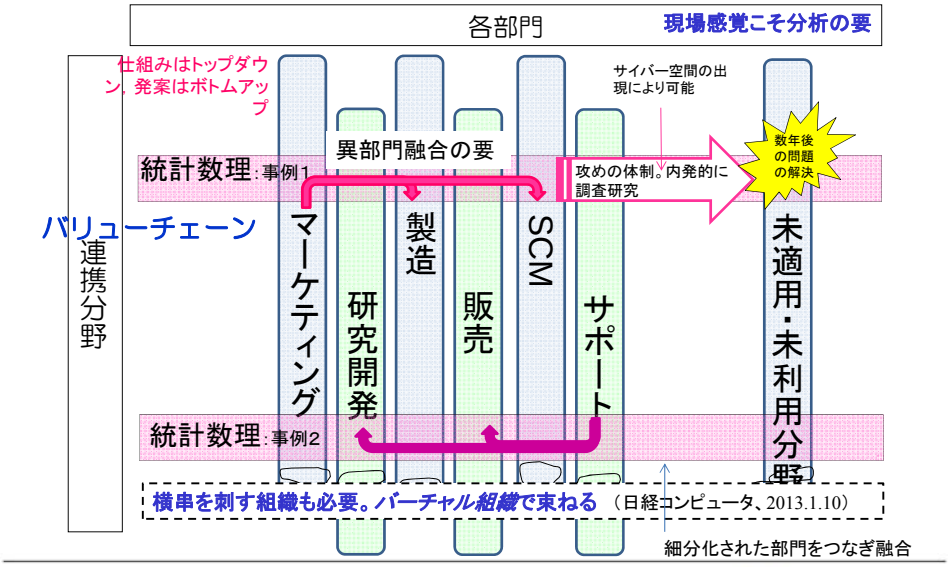
発見科学研究 成果報告集
宮野悟(平成13年3月)

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

6. 対応策：人材育成

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

企業を横断する中核部隊として

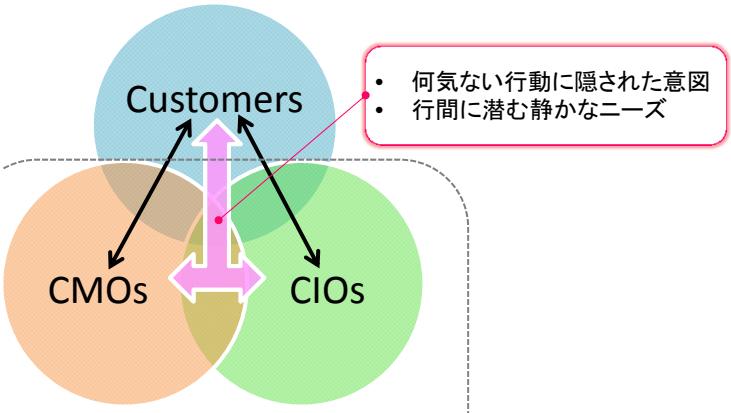


統計数理研究所

CMOとＣＩＯの協働作業が大切

CMO: Chief Marketing Officer CIO: Chief Information

データの価値次第でインフラやアーキテクチャの設計が大きく変わってくる。



http://blogs.hbr.org/cs/2012/11/why_cmos_and_cios_need_to_team.html

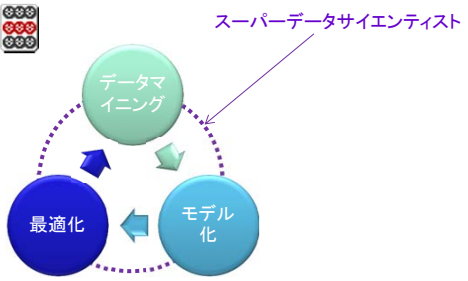
統計数理研究所

データ解析・分析の『個人商店時代』の終焉

•ちゃんとした、横断的チームが必要

■理論，モデリング，計算，実装，そして応用分野（現場）の専門家との**協調（協働）**作業

■**一気通貫**方式で知識発見プロセスを指導できる**リーダー**



Data Scientists are actually T-shaped

5 points for job description

- ✓ Innovative problem solvers who learn from data
- ✓ Expertise in statistical modeling and machine learning
- ✓ Solid understanding of problem domain
- ✓ Effective communicators of what they learn
- ✓ Specialized programming skills

（By R.N. Rodriguez）

Data Businessperson: Determining benefit of data projects to organization
Data Creative: Applying broad range of analytics and technology
Data Developer: Acquiring, storing, cleaning, and managing data
Data Researcher: Understanding complex processes

Data Scientists tend to have deep experience in one category and some ability in others.

46%の会社がデータサイエンティストの雇用を増やす予定 （2011 Bloomberg Bussinessweek Survey, 930社回答）

2018年までにデータサイエンティストは16万人不足。
米国では毎年4千人育成。500人（博士）、2400（修士）、1100（学士）。修士と学士は急増中。

US: Data Scientist　4つのタイプ

Binita

Data Businesspeople
- 経営工学MBA
- コンサルティング会社での経験
- 現在大[illegible]業のデータ分析
[illegible]通
[illegible]いが、
[illegible]る

ソフトウェア開発ができない！

Rebecca

Data Researcher
- 分子生物学で博士号を取得
- 元々はアカデミア指向
- 現在は国際的な流通会社でデータサイエンティスト
- 論文は書けるが、マネジメントの経験はない

Chao

Data Creative
- 経済・CS・統計
- 統計コンサルのベンチャーを起業
- 現在大手新聞社に勤務
- 夜はPythonのオープンソース開発
- 自身はハッカーと思っている

Dmitri

Data Developer
- CS修士
- 現在[illegible]ンサルファー
[illegible]ルの開発
[illegible]が書け

ビジネス改革ができない！

（丸山教授@統数研　のスライドを改編）

INSIGHT DATA SCIENCE
FELLOWS PROGRAM

Who makes the best data scientists?

Who makes th[illegible]er Chief
Scientist at Linked[illegible]nsight
mentor, "the best [illegible]
physicists, rather t[illegible]g
mathematical bacl[illegible]in which
survival depends c[illegible]bout the
big picture, the big[illegible]work is
highly quantitative[illegible]onomy,
astrophysics, phys[illegible]s well as
researchers in mat[illegible]perations
research, economi[illegible]ields.

Insight Datascience Fellows Program

カナダ人。アカデミアのキャリアから「ドロップアウト」してシリンバレーに 来る。いくつか企業

Ph.D 量子物理

Ph.D 物理

Ph.D 物理

丸山、John、Jake、Cathy

・2年間で、およそ100名のフェローを輩出
・全員就職

・プロジェクトの内容は、各フェローが自分でアイディアを出して決める。
・すべて公開されているデータを使う。

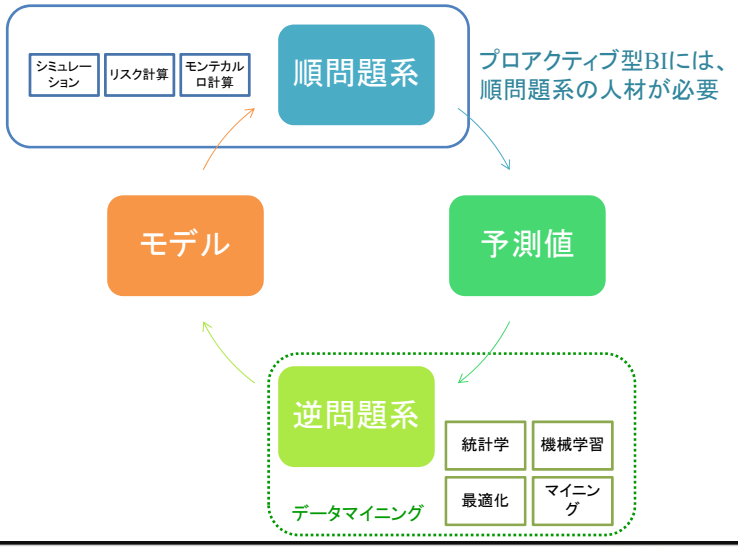
・最新のセッションでは、500名の応募があった（すべてPh.DまたはPh.D candidate）。
・採用試験は電話によるインタビューと、プログラミング・統計のスキルを問う。

・Web広報以外には、大学、特にその就職支援部門へ行ってこのプログラムの紹介をしている。

・夏、New Yorkに新しいオフィスを出す。
・現在はPh.Dに限っている「サイエンティスト」のプログラムを、エンジニアにも拡大予定。

（丸山教授@統数研　談）

プロアクティブ型ＢＩと人材



日本: Data Scientist　４つのタイプ

メーカーの製品開発・企画部門にいる中堅のIT系エンジニア。
社内では確実にデータの活用が進んでいる。キャリアパスも見えている。

主に中小のサービス系の企業に勤める女性。
比較的自由になる勤務形態を望んでいる。

若手で、まだ実務経験は少ないが、データサイエンティストになりたい夢を持っている。

ITサービス業でデータ分析をプロとして長年実施してきていて、この仕事に誇りを持っている。

（丸山教授@統数研　のスライドを改編）

ビッグデータ時代における
データサイエンティストの役割

仕組みを生み出す。

- 常識的（合理的）判断を自動的かつ超高速に実行
  - ✓ 外れ値、異常値の発見→イノベーションへ
- ひらめきを得る作業のアシスト
  - ✓ ボナンザ：力づく探索、ボナンザ囲い

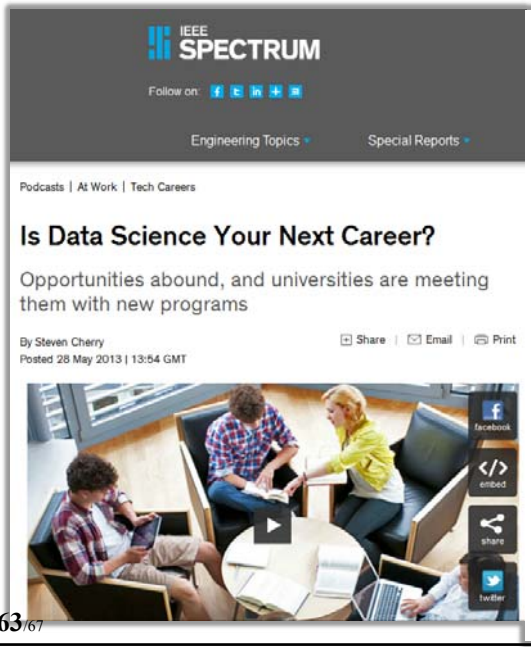
歩		歩		
	歩	銀	歩	歩
		玉	金	金
香	桂			

ボナンザ囲い

- 異分野融合の要
  - ✓ 近年の先端テクノロジーは異分野融合が主
  - ✓ ビジネスモデルの創発エンジン
- データをガイドとして二元論的思考から循環論的思考へ
- データマイニングは列挙と探索空間の枝刈りが基本。思考法がちょっと異なる。いわゆる、順問題系。

Deep Analytical Talent

MGIリポートから



IEEE SPECTRUM

Follow on:

Engineering Topics | Special Reports

Podcasts | At Work | Tech Careers

Is Data Science Your Next Career?

Opportunities abound, and universities are meeting them with new programs

By Steven Cherry
Posted 28 May 2013 | 13:54 GMT

Share | Email | Print

Steven Cherry: Hi, this is Steven Cherry for IEEE Spectrum's "Techwise Conversations."

In a recent podcast, I was surprised to learn **there were 93 000 data scientists registered with Kaggle,** the site that creates competitions among them and helps award freelance contracts. I'm not the only one. The article in The Atlantic that brought Kaggle to our attention had a parenthetical exclamation: "Who knew there were that many data scientists in the world!"

The next obvious question is, How do I get myself in on the lucrative area of data science? As it happens, the trusty New York Times wrote about that back in April. Its premise was, **universities offer courses in a hot new field, data science. Data science sure is hot**. Is it new? Sort of.

The Times article quoted an adjunct professor at **Columbia University**, who described a data scientist as **"a hybrid computer scientist/software engineer/statistician."** And it says that this fall, "Columbia will offer new master's and certificate programs heavy on data. **The University of San Francisco** will soon graduate its charter class of students with a master's in analytics. Other institutions teaching data science [JPEG] include **New York University, Stanford, Northwestern, George Mason, Syracuse, UC Irvine, and Indiana University**."

My guest today, also with Columbia University, is Chris Wiggins, a professor of applied mathematics there. He'll walk us through what's old and new in the academic field of data science. He joins us by phone.

Chris, welcome to the podcast

やっぱりだめか？

DataScientist Society

一般社団法人
データサイエンティスト協会

幹事会員

SAS Institute Japan
電通
日本アイ・ビー・エム
博報堂DYメディアパートナーズ
ブレインパッド
ブロードバンドタワー
ヤフー

賛助会員

アイウェイズ
ALBERT
オープンストリーム
オリコム
ジーリサーチ
損害保険ジャパン
デジタル・アドバタイジング・コンソーシアム
博報堂プロダクツ
日立インフォメーションアカデミー
Pivotalジャパン
フュージョン
三菱電機インフォメーションシステムズ

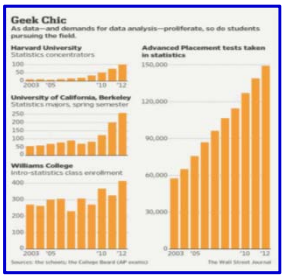
＊2014年4月現在、会社名50音順

諸外国の統計学科（米国編）

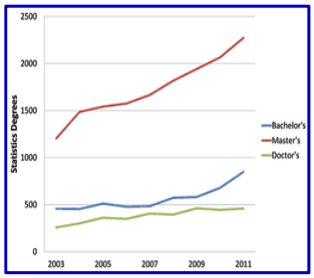
➢ 有名私立大学筆頭

Stanford University
⇒Department of Statistics
約30名の教員が在籍

男性教員：約27名
女性教員：約 3名
男女比率：約10%



米国において高まる学生の統計学への関心
（Wall Street Journal.）



米国における統計学の学士・修士・博士号取得者数の推移（2003-2011）
http://magazine.amstat.org/blog/2013/05/01/stats-degrees/

➢ その他の有名私立大学

Johns Hopkins University, Washington University, Duke University,
Boston University, Purdue University, Carnegie Mellon University, Rice University,
Texas A&M University, Georgia Institute of Technology, Temple University等
⇒Statistics関連の学科が存在している

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

諸外国の統計学科（アジア編：中国）

➢ 中国＝2000年以降，毎年大学における統計学科の数を増加

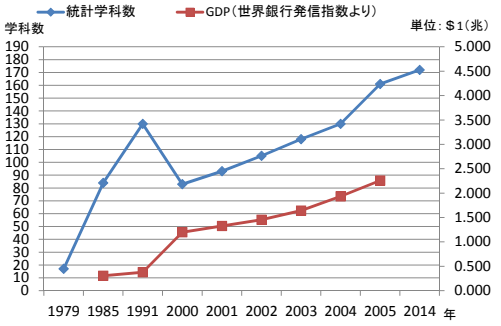
＜統計を専攻する学科数の推移＞

1979年：　１７学科

約四半世紀後

2005年：**１６１学科へ増加**

2014年：**１７２学科へ増加**



年度	1979	1985	1991	2000	2001	2002	2003	2004	2005	2014
統計学科数	17	84	130	83	93	105	118	130	161	172

＊2005年までの数字は日本学術会議報告「数理科学分野における統計科学教育・研究の今日的役割とその推進の必要性」より引用
＊2014年の統計学科数の数字は中国の教育部（http://www.moe.gov.cn）から指定された大学入試情報ホームページの検索サイト（http://gaokao.chsi.cn/sch/search--ss--on,option-qg,searchType-1.dhtml）より引用