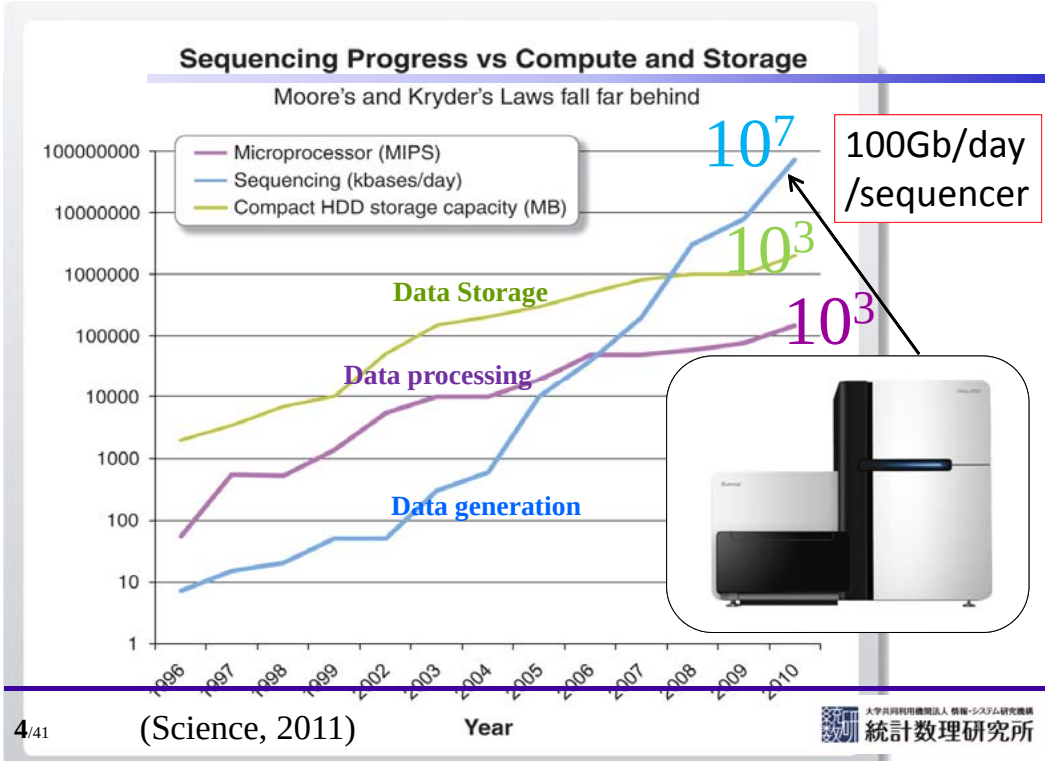
下流からせめるビッグデータ

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



(Science, 2011)

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)

ビッグデータがなぜ大切か？

・生活を“まるごと”とらえた結果
・支配方程式のない現象も科学へ

ビッグデータ環境下における研究開発推進の鍵

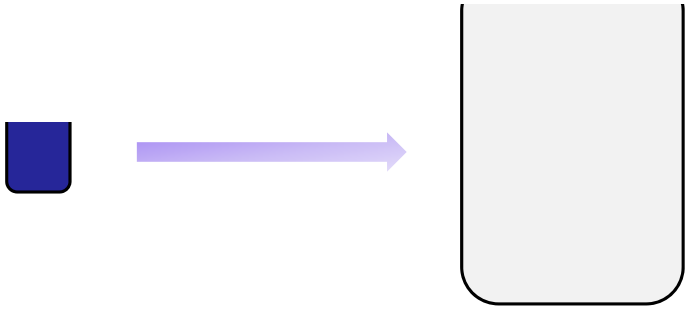
ビッグデータは

「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

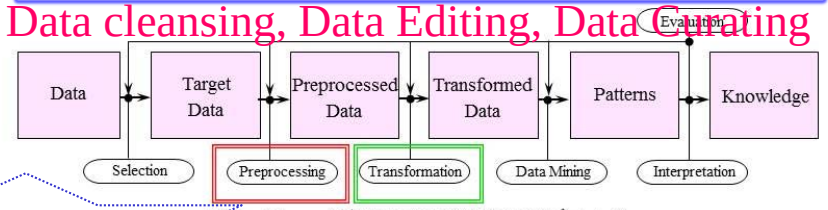
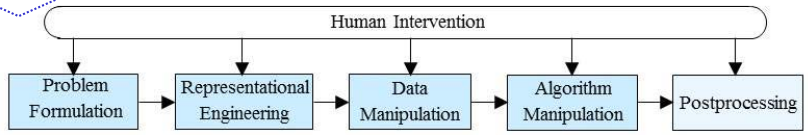


図 1: Fayyad等による知識発見のプロセス

13年前の資料から



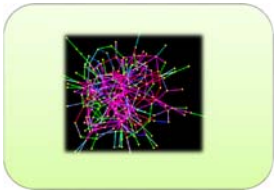
データ処理の流れ 図 2: **Langley**による知識発見のプロセス

発見科学研究 成果報告集
[illegible]

5. 日本固有の問題点

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

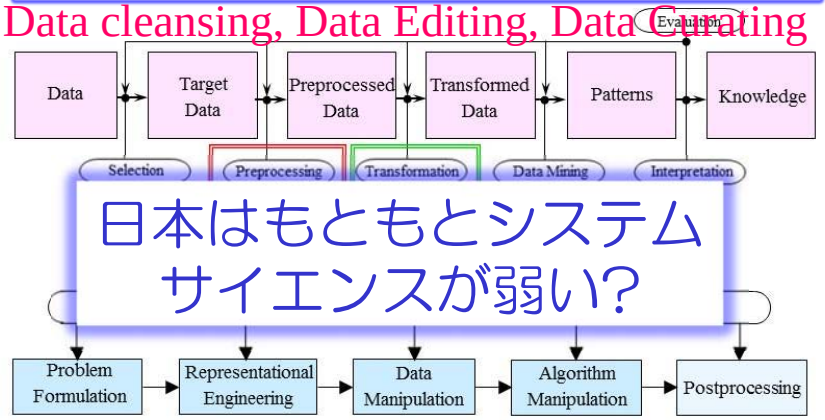
「もの」は分解できる

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



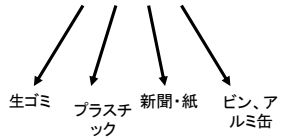
日本はもともとシステム
サイエンスが弱い?

データ処理の流れ 図 2: **Langley**による知識発見のプロセス

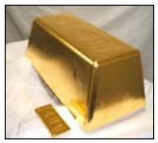
発見科学研究 成果報告集
[illegible]（平成13年3月）

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
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

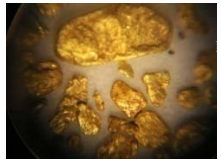


分別、整理することで



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



マネーボールの実在

玄人の「眼」，「聴」を造る

世界に誇る東京のモノづくり
輝く技術 光る企業 kirari-tech
http://kirari-tech.metro.tokyo.jp/

いつまでも「プロジェクトX」の感傷に
ひたっていては世界から取り残される

機械は許しても、目が、指先がわずかな誤差を許さない
顧客のニーズが厳しくなればなるほど、経験に裏打ちされたその技術が冴え渡ります。

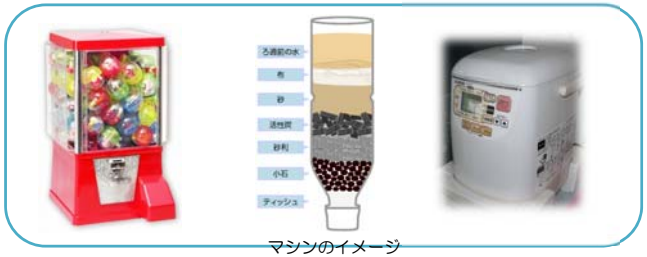
『マシンに入れれば何かでてくる』は幻想

■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

6. 対応策：人材育成

個人情報保護とデータの本格的利活用

One for All, All for One

One for All,

ひとりはみんなのために、みんなは日本のために。

All for Japan.

「IDデータコモンズ」はドナーカードのようなもの

（情報研　曽根原教授）（日経ビジネス2013.09.30号）

帰納と演繹

技術（客観的、絶対的） → 価値（主観的、相対的）

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

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！

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」

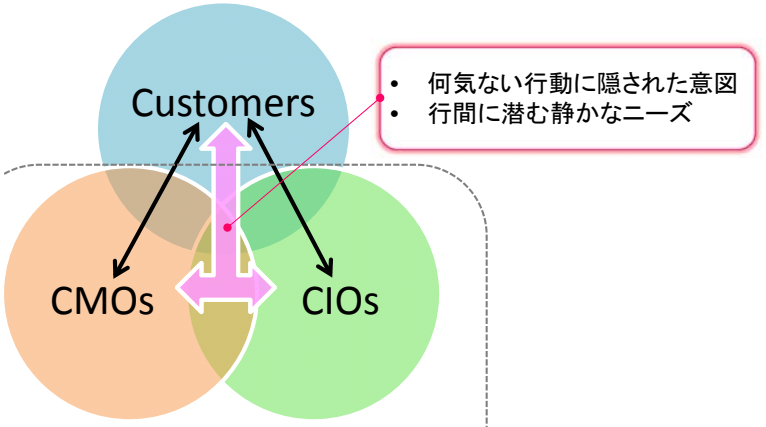
■非常時： Forecasting機能が大切

- 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

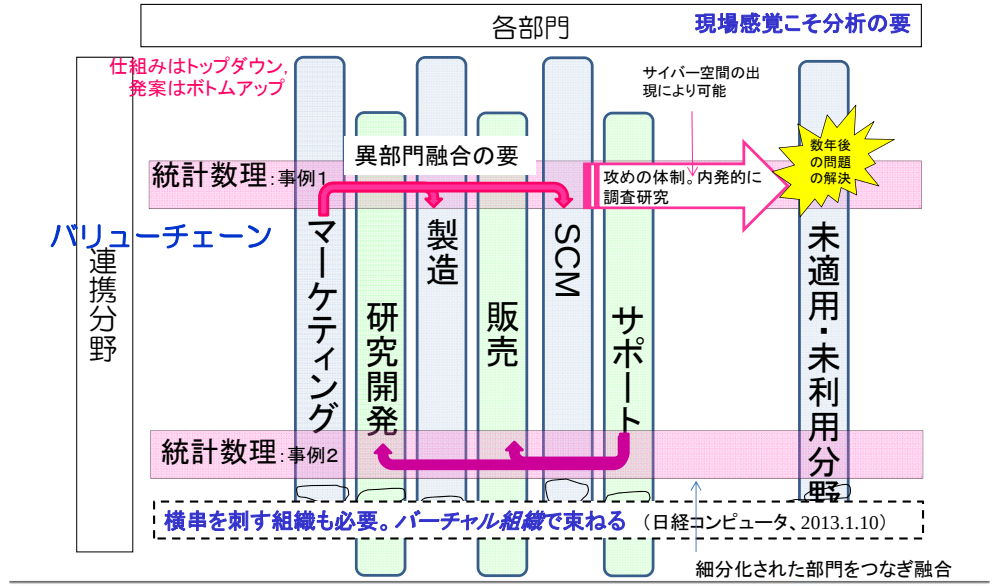
CMOとＣＩＯの協働作業が大切

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。

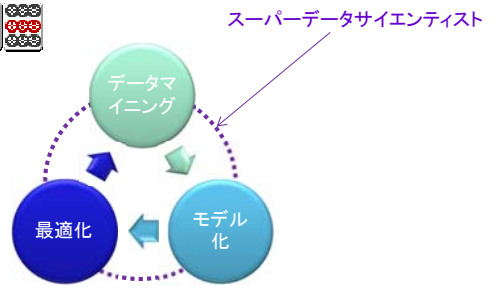


企業を横断する中核部隊として



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**



Insight Datascience Fellows Program

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）

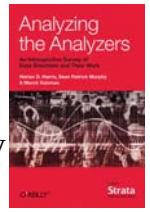
Data Scientists are actually T-shaped

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]

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 Creatives
- 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]堅[illegible]ンサルファー[illegible]ルの開発[illegible]が書け[illegible]

ビジネス改革ができない！

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。

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

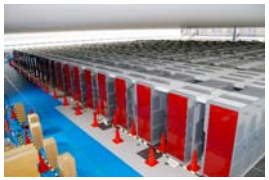
$$\max . f(\theta) \quad \theta' = (\theta_1, \ldots, \theta_p)$$

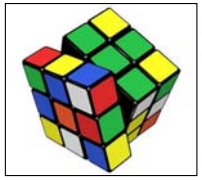
パラメータ数が2個（p=2）なら、10 x 10 =100 通り計算すればよい。

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。**

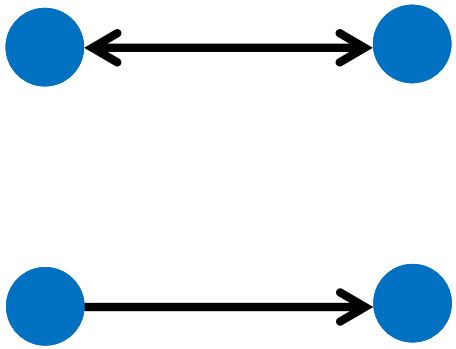
率（頻度，指標）がくせもの

$$\text{リピート購買率} = \frac{a}{A}$$

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

落とし穴2：列挙処理、相関と因果



Granger causality

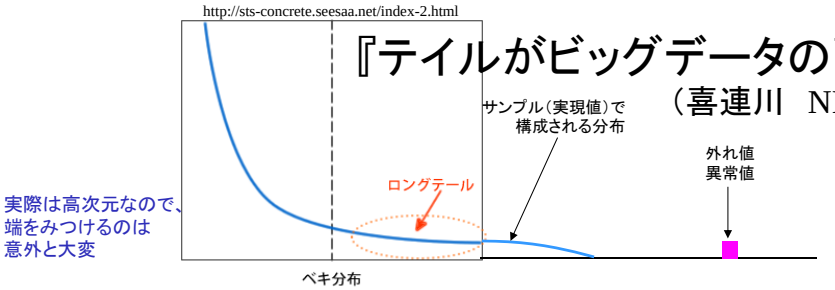
2003 Nobel Memorial Prize in Economic Sciences (from Wikipedia)

- 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』

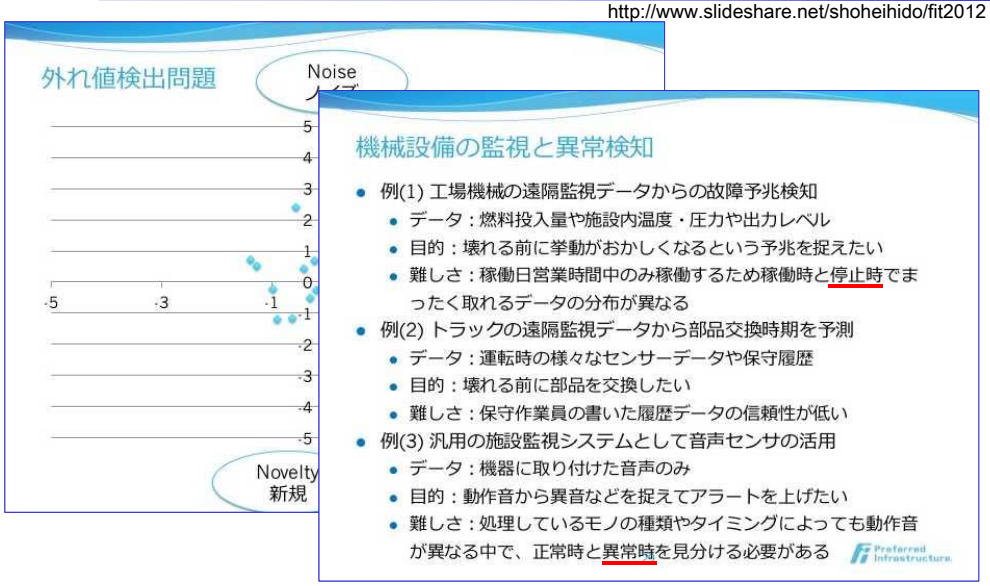
- 新発見、ひらめき
- クレーム（PL法対応）

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

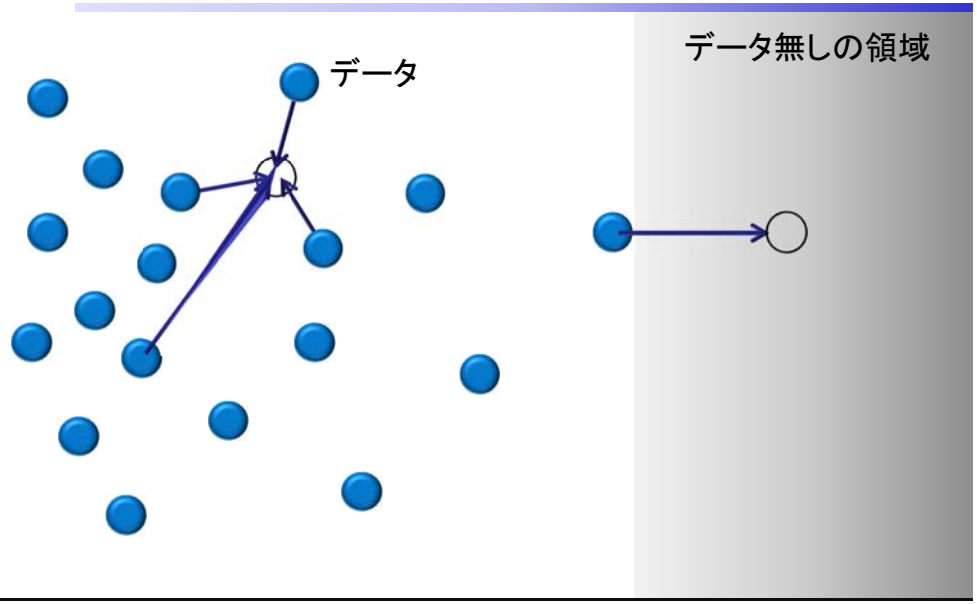
外れ値の多様な呼称（PFI 比土：Jubatus TL）

http://www.slideshare.net/shoheihido/fit2012



統計数理研究所

落とし穴4：内挿と外挿問題

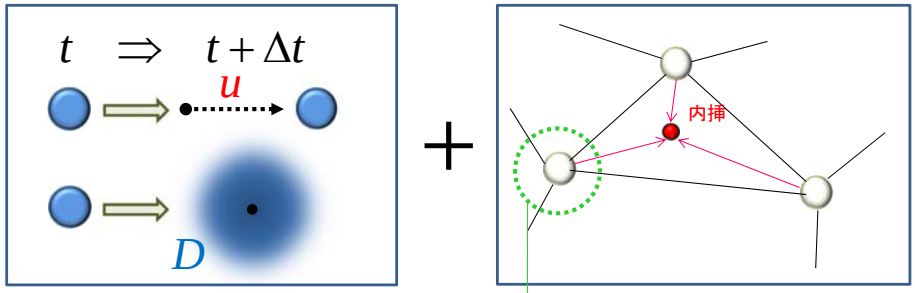


統計数理研究所

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）