
ISP Networks 設立記念セミナー
講演３
「ビッグデータにとりかかる前に」宇野毅明氏

２０１５年６月２９日（月）
ISP Networks 設立記念セミナー　式次第

14:00　開催のご挨拶　代表：**佐々木賢二**

記念講演

14:15　「NEDOのCPS/IoT」 山崎光浩氏
14:45　「データ活用ビジネス最前線」 東富彦氏

15:55 – 16:05 休憩

16:05　「ビッグデータにとりかかる前に」 宇野毅明氏
17:00　まとめ

17:15 – 20:00 **Networking Party** （隣室）

[講演3] 16:05 – 17:00

「ビッグデータにとりかかる前に」

イノベーションの創出や価値創造、コスト軽減や単なる趣味まで、ビッグデータの解析と利用は大変注目を浴びている。データの研究や活用には既に様々な研究が行われてきたが、データの分析・活用の方法にはこれといった万能的な処方箋がなく、他の事例との類似性も薄いため、参考とすべき解析技術、手法があまりない状態にとどまっている。利用場面ごとの類型化も難しく、課題設定から技術の探索まで、全てを一から始めなければいけないのが現況である。今回 は、ビッグデータに対する「感覚」を講義したい。最近のビッグデータとはどのようなもので、どんな感じで使うとどのようなことが得られるのか。既存技術の 位置づけや注目すべきポイントなどを、「コツ」を含めた形でお伝えしたい。


[講演3] 16:05 – 17:00
「ビッグデータにとりかかる前に」
講師：宇野 毅明（うの たけあき）氏
大学共同利用機関法人　情報･システム研究機構
国立情報学研究所
情報学プリンシプル研究系　教授　理学博士

ビッグデータにとりかかる前に

宇野 毅明　（国立情報学研究所
総合研究大学院大学
JST CREST ビッグデータ領域）

http://research.nii.ac.jp/~uno/index-j.html
e-mail: uno@nii.ac.jp

2015年6月29日　ISP Networks 設立シンポジウム

国立情報学研究所　宇野 毅明

東工大、情報科学卒：東工大経営工学 → 現職

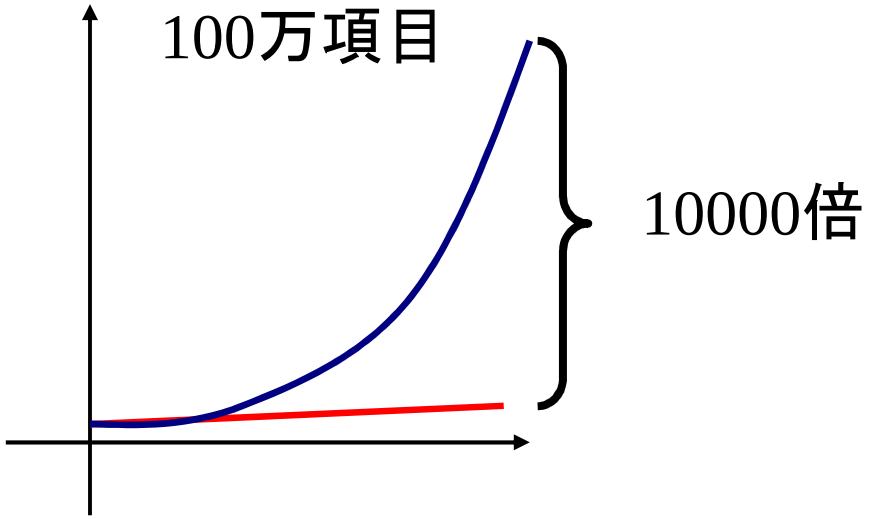
専門分野：アルゴリズム理論　（計算量＆実利用）

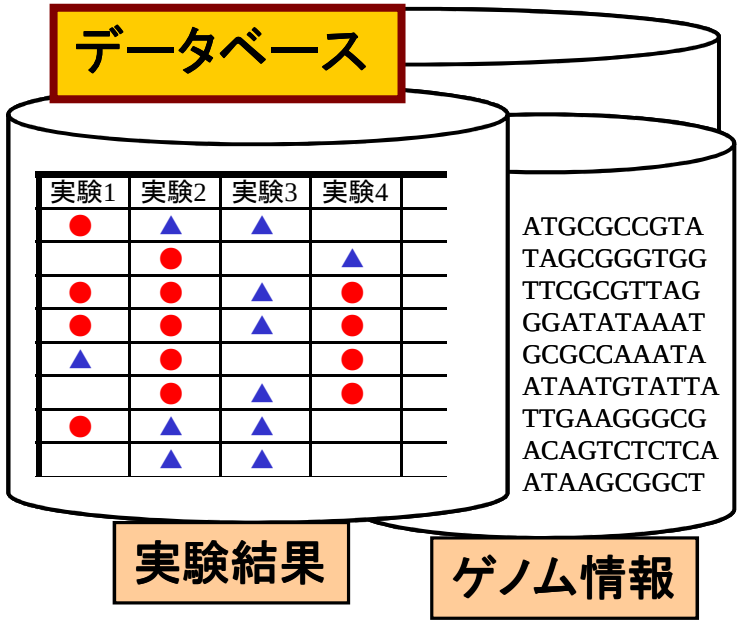
計算方法の改良による高速化の研究。

データ規模増加に対する計算時間の増加カーブを改善

大規模データに対する、基礎的な情報処理に取り組む

頻出パターン、類似性解析、クラスタリング、可視化、最短路・・・、などに使われる基礎計算





データビジネスの課題

ビジネスとして何が大事か

- ビジネスにおける、新しく考え出す必要のある物事の（概念的な）重要度を考える

ノウハウがない、やり方で大きな差が出る、という意味では、発想と判断のところに大きなウェイトがあるだろう

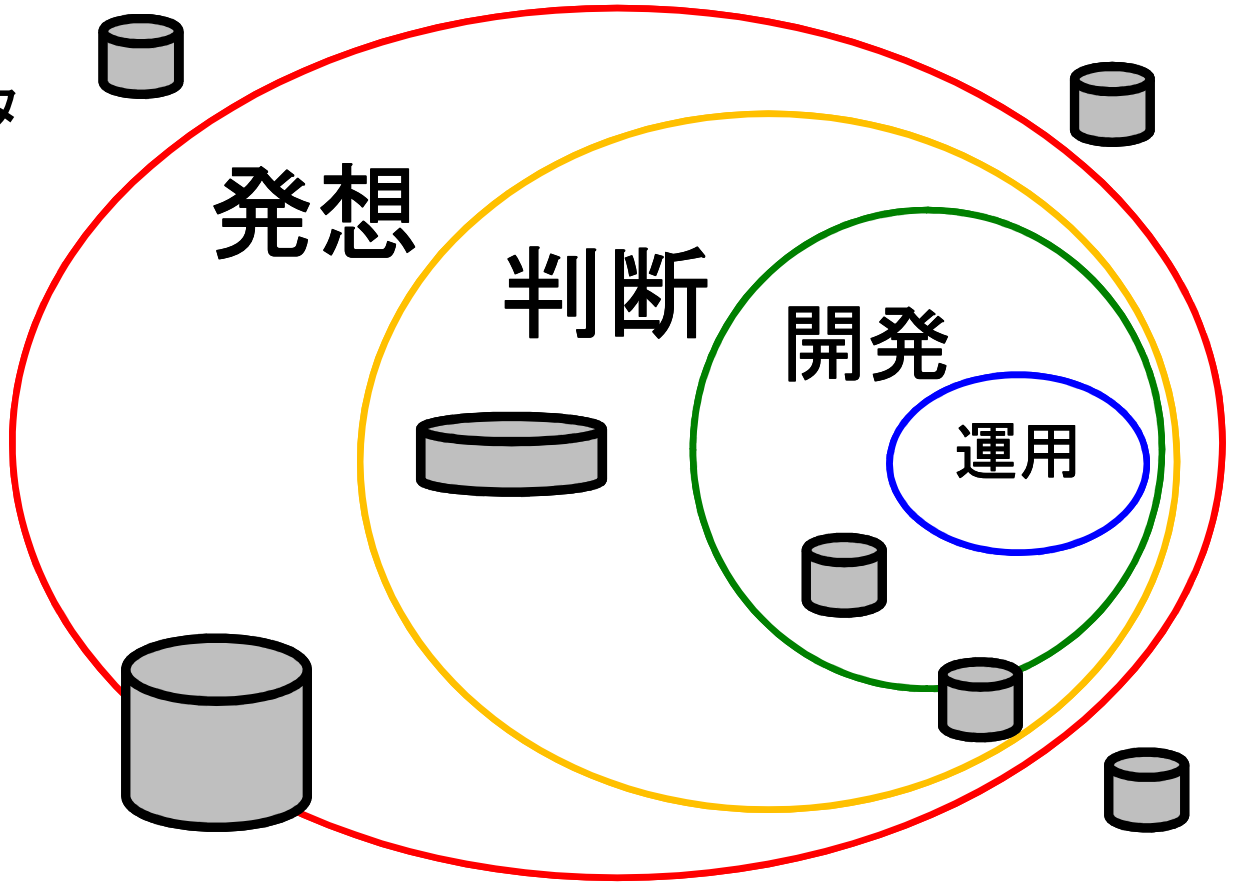
そういう意味で、ビジネスの上流段階で、データは重要

（発想の種、判断材料）

目的の抽象度

- 上流のことは概念的に大きく、現場に行くに従って、目的は具体的になる
  → データは、自分の欲しいところにあるとは限らない

- 自分の目的とデータから得られるものを、どう結びつけるか、それを適切な抽象度で行うのが大事



例えば

日本の全ての電柱にセンサをつけて「音量」を測る
このデータで何ができるか、考えよう

先ほどの分類で整理してみましょう。

（発想、判断、開発、運用）

例えば

運用

+ 倉庫入口の電柱の音量から、トラックの出入りを自動収集
+ イベント会場のそばで大音量を出してないか
+ 夜中の町の騒音と、静寂故の危険を調査
+ ゲリラ豪雨の探知
+ 夜中の公園や河原の監視

開発

+ 防犯ニーズの調査、ニーズの高そうな地域の選定
+ ゲリラ豪雨予測システムの検証
+ 社内にもマイクを置き、新型携帯デバイスの騒音の調査
+ 工場内スケジューリングシステムの検証

例えば

判断

＋ 店舗出店にあたり、出店エリアの、時間別の賑わいの調査
＋ イベント会場や観光地のシーズンごとの賑わいを調査
＋ 通学路・最寄り駅への通勤経路となる道の把握
＋ 祭りの警備の計画

発想

＋ 町の音が出ている部分を可視化
＋ 町の、賑わいと人の流れの移り変わりを可視化
＋ 住宅地の静けさ度合い地図
＋ 各地域の、経済活動の活発さや、ライフスタイルの分析
＋ 音を出すデバイスによって、自分の位置を知らせる

ここでの考え方

（鬱病） 元気がない人を見つけたい

← 元気がない人は、忙しい・遅刻がある、部局が大きく変わる

自分の見たいものを、他のもので代用している

（会議） 会議回数を元に社内環境の整備

← 実際に使われるものを用意する

自分が計画を立てる際に、判断材料にしている

（人財） 社内の人財を有効に活用したい

← 余力のある人を見つけ出し、切羽詰まったチームに送る

人財活用という大目標を、具体的で細かい物に落とし込む

結びつければいくらでも

＋ 社内で購入した事務用品・PCから、資産の活用方法を検討

＋ 勤怠管理、購入、会議録などの作業データから、「うまい仕事の仕方をしているチーム」を抽出

＋ トラブル対応の回数で、保守管理をしている資源の質の評価

＋ 材料・スタッフの違いによる歩留まりの評価

…

過去の実例

- データを活用した企業の成功例・失敗例は数多くある

しかし、その多くは「本質的な部分」の解説がない

我が社は〇〇で集めた△△のデータを最新の□□システム
を用いることで、効率を20%上げました
（この化粧品は天然〇〇100%ですので、お肌つるつるです）

- 本当に知りたいのはここではない。こうなっていて欲しい

△△のデータは〇〇の要素を含んでいるので、□□の情報が
得られるはずです。が、既存システムでは××の理由により
それができなかったので、新しく…

- 事例を出した人も、本質はわかっていないのでは？

歴史の短さ

- 歴史の長い分野では、物事のメカニズムがよく分かっている
ので、技術や物事の解説に確かさがある

この車は、レーダー測距によって、障害物があると自動的に
停止します。だから事故が減ります。

- この「既存業種では当たり前」が、データ産業には通じない
  理由は、獲得してきたノウハウが少ないから

- これから、みんなで作っていく物なのでしょう
  逆に言えば、誰でも最前線です

やるべきこと

- 歴史が長ければ、やるべきことはすぐ分かる
  勉強して、技術を教えてもらう／買う、とすればいい

- しかし、**黎明期の分野は、何を知るべきかも分からない**
  逆に、何をやっても最前線だし、簡単なもので成果が出る

- こういう状況、特に情報の分野では、**「意見交換」が最も大事**

データを持つ人、知る人、現場の人、解析技術、ビジネスプラン…

… つまり、こういう人が集まって飲み会をすればいいと言うこと
  当たり前ですが、実際、できていないですよね？

情報・意見の交流

- 実際に、私も多くの研究者・企業人と話をする

- 企業・学会を問わず「目的意識と課題の明確化」が
  議論の中心になることが多い

- 日本人は、知識を話すのは上手だが、意見を言うのは苦手
  なので、打合せや学会では、「いい結果が出ないことが多い」

- 飲み会では、自分の意見を主観的に言うので、話が進む
  そういうことを最初から狙って、ワークショップを行ってもいい

データ解析の俯瞰

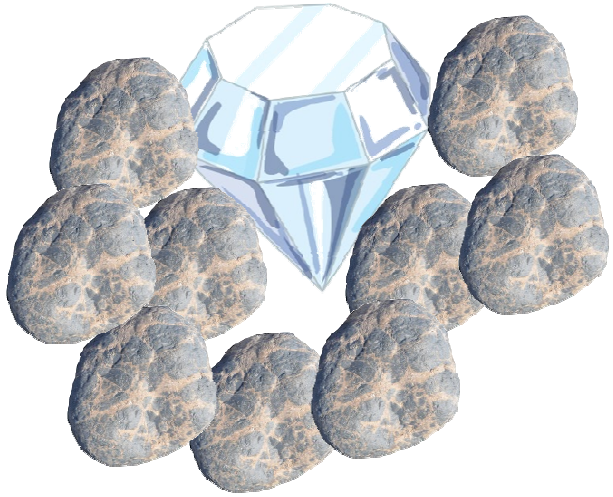
バズワードとしてのマイニング

- バズワードとしてのマイニングは、ずばり「データ解析」

- 他にも似たようなバスワードがたくさん
  IT化、電子化、Web2.0、クラウド、ビッグデータ、見える化…

- バスワードは「宝石にゴミを付ける方法」
  つまらないものを、良く見せたい

- こういうときは、「正体を知る」ことが大事
  データ解析を俯瞰しましょう



データ解析とは

- ものの理屈は、自分の経験や知識から導き出すもの
  ＋ うちにはこういう客が来るだろう
  ＋ 今週末はこういうことがあるから、客足が遠くなるだろう
  ＋ なんか客層が変化してるけど、時代がこう変わったからかな

- 対してデータ解析は、データを見ることで、知見を得ること
  ＋ 最近年配の客が増えている。理由は何だろう
  ＋ 景気の変動とうちの客層、関係があるかな
  ＋ うちの客の家を地図に書いてみよう。どうなってるかな

- いろいろなことがわかる。そこから始まるアクションは2つ

  作戦1 理由を考える： なんで最近この商品が売れているんだろう？ ああそうか、それならこういう感じで売ってみよう

  作戦2 条件反射： この商品が売れてるな、たくさん仕入れよう

もっと高度なことを

- 単純な特徴は、理由を理解しやすい
  それに対する作戦立案も、立てやすい
  しかし、洗練されたビジネスでは、それでは不十分

- 「なぜこういうことがおきるのか」がわからないと、ビジネスが設計できない

  ＋ 今日起きたことが明日起きるとは限らない
  ＋「仕組み」がわかれば、その変化と変化の質は予想可能
  ＋ 特徴的な事実の全体での位置や、他者との関係性も大きな情報

- こういったことを助けるのが、いろいろな情報技術

解析技術の大まかな分類

- 「手法」と「データの種類」と「インフラ」があることに注意

クラウド

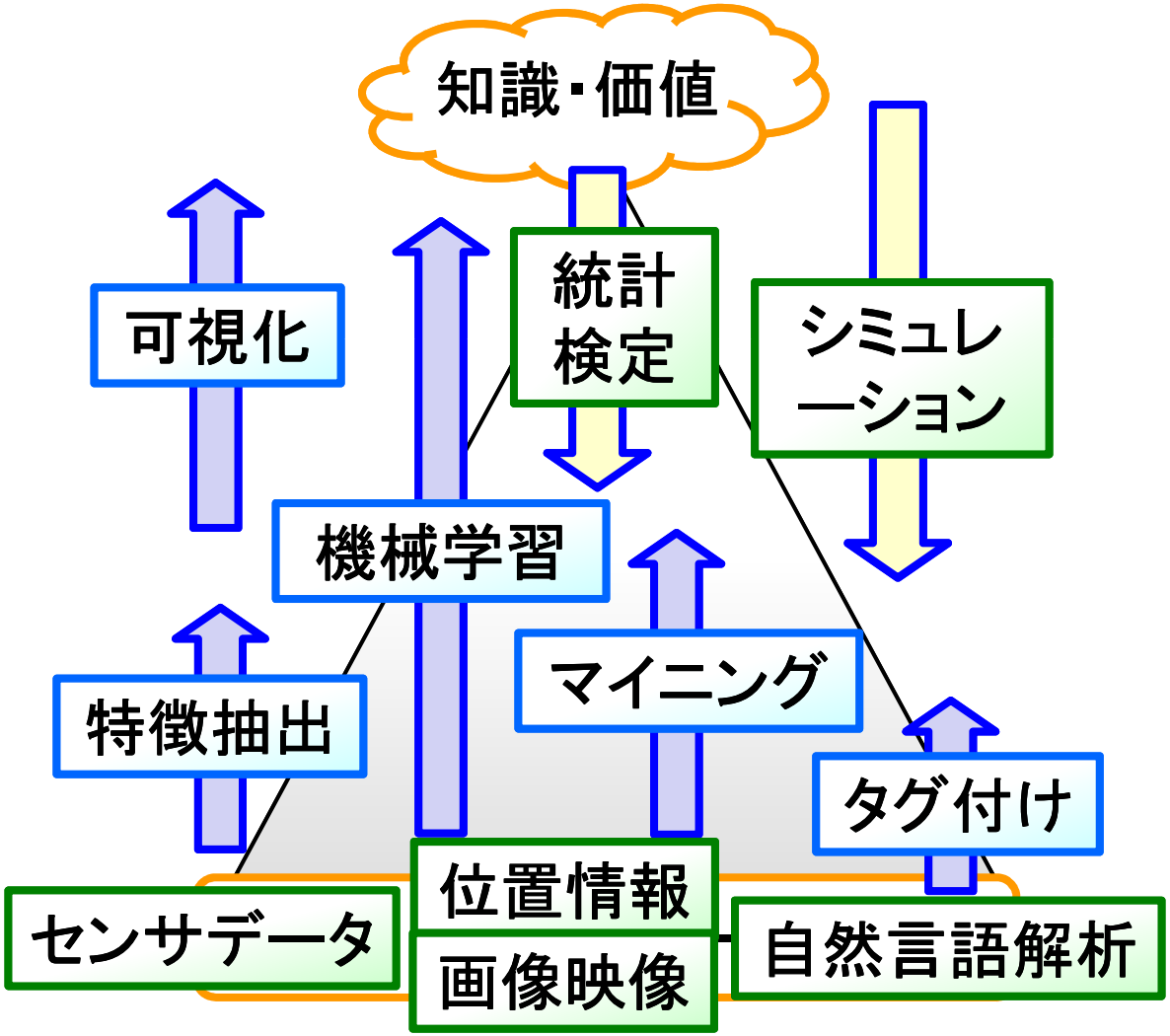
センサ

プライバシ

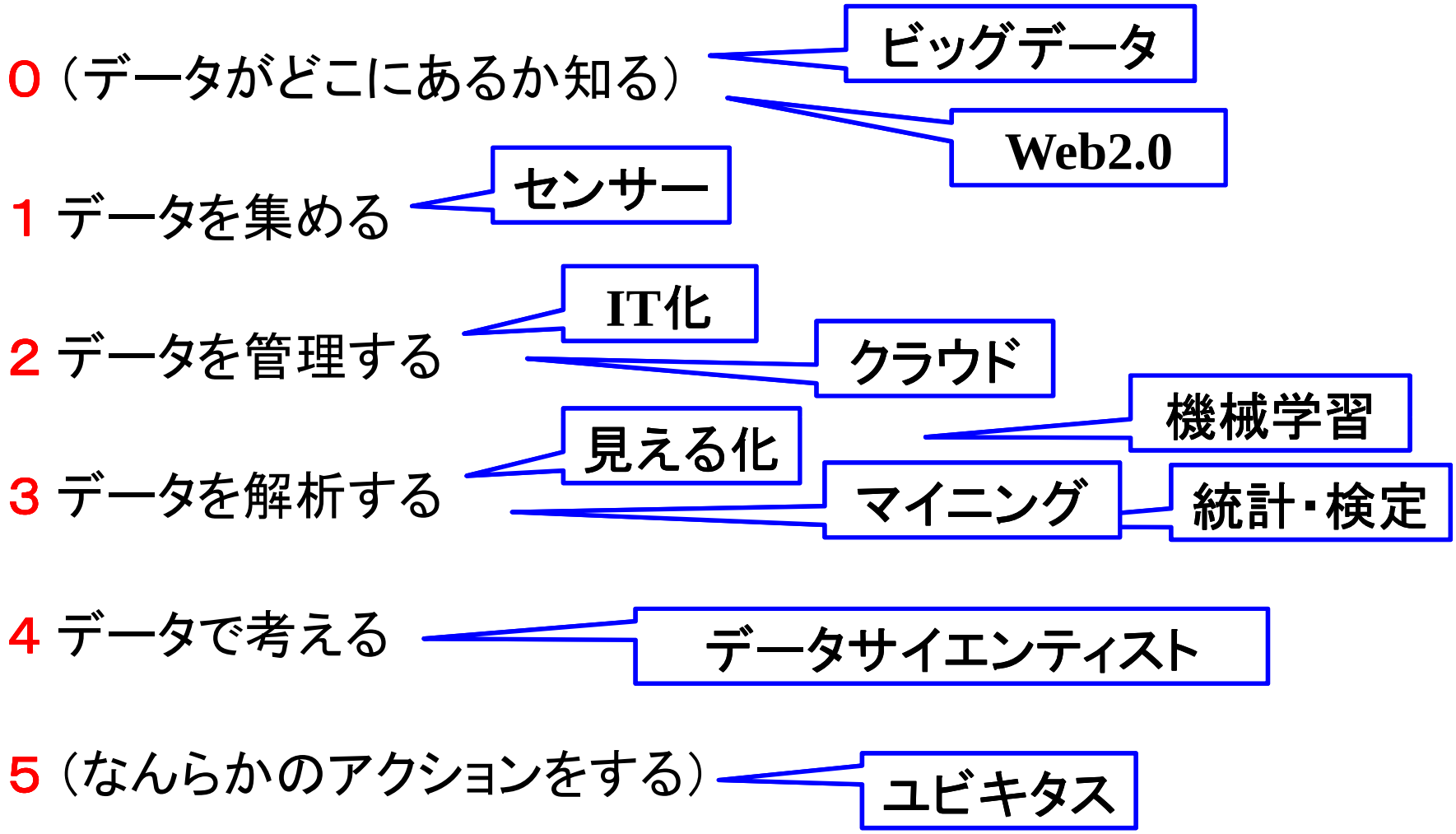
アルゴリズム

ゲーミフィケーション

クラウドソーシング



データ解析の流れ



一部に革新があると、それを唱えて全部を売り込む

統計・検定

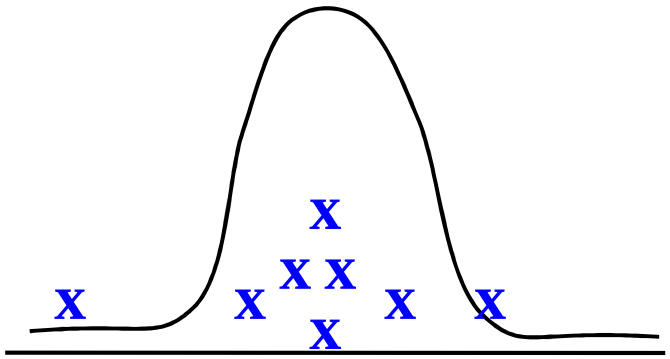
- データはランダムに、ある種の「分布」← モデル にしたがって発生する、という原理原則に基づいた解析

- 観測したデータが分布に従って発生したのならば、その分布の中心はどこで、どの程度の大きさなのかを推定する

- この「推測（当てはめ）」を使って、
  検定、異常検出、区別、などを行う

- 「全体的な傾向」を見るのが得意

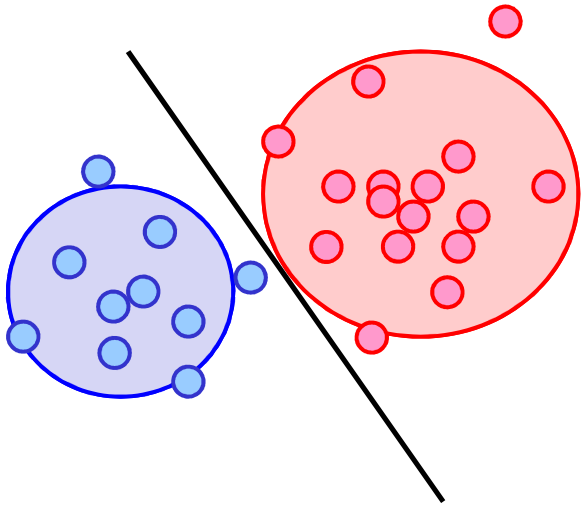
マイノリティや、多様性は苦手



機械学習

- 統計的な手法と計算を使って、より知的な判断を行う手法

＋ クラスタリング
＋ 属性推定
＋ 特徴選択
＋ 異常検出
…



- 統計より「**より高度なタスク**」ができる

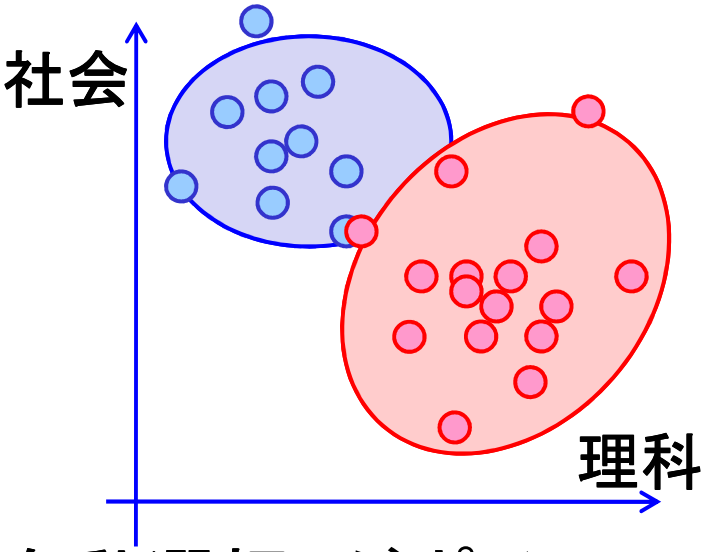
- 「大きな、全体的な操作」をする、演繹的な手法

マイノリティや、多様性は多少扱えるが、やはり苦手

理系／文系の推定

- 社会と理科のテストの点数から、その人の志望を当てる

＋ 理系／文系の人のテストの
点数データを集め、分布を見る
＋ 分布が分かると、両者を比べて
「ここは理系っぽいかな」がわかる
＋ その数値を元に予測する



- 「見るべき変数とその組合せの自動選択」がポイント

- 理系文系の「分布が分かる」
- テストの点から、「文系理系が推測できる」
- ちょっとおかしな志望をしている人が分かる（異常検知）

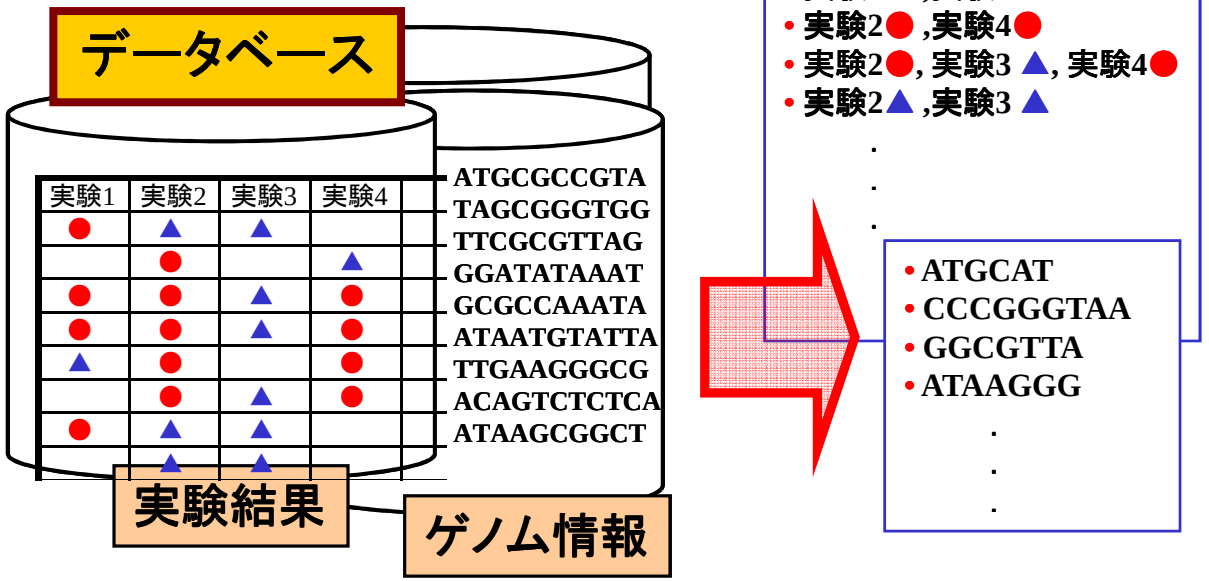
データマイニング

- データベースの中から「面白いもの」を見つける
  - 面白い ⇔ 沢山現れる、他と似てない、集中してる…
  - もの ⇔ パターン、グループ、セグメント、項目…

- 「データ帰納的」「発見的」

何かの知識を直接的に得るものではない

＋ 網羅性がある
＋ 多量の解が出る
＋ 自動的なプロセスではない
＋ 意味解釈、利用に難かしさがある



伸びる科目は？

- テストの点から、数学が伸びる／苦手になるシグナルを見つける
  ＋ 個人の成績データを収集
  ＋各人の成績の中から「シグナルになりそうなもの」を選び出す
  ← ○○の成績が伸びた／落ちた、など
  ＋頻出する「シグナルの組合せ」を見つける
  単に頻出、苦手になった人に頻出、伸びた人に頻出、、、

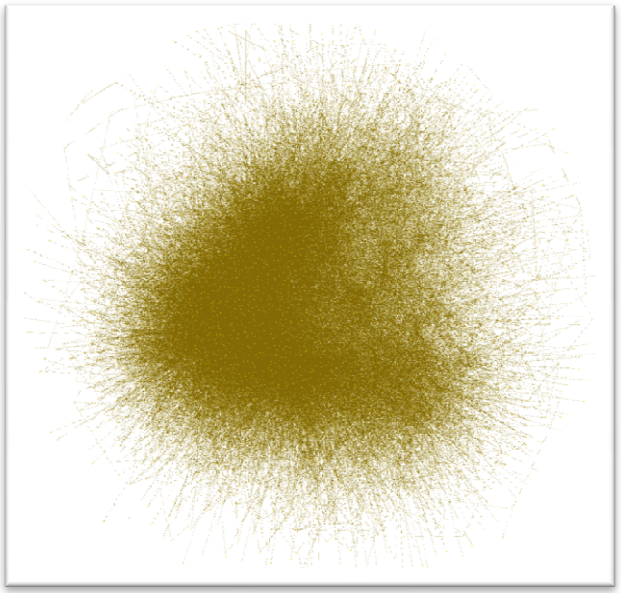
- 「シグナル」は見つかる
  自明な物が多いだろうが、
  面白い物もあるかも

- 数理的に何かを言うのは難しいが
  数理的に想定外まで見つかるかも

	1学期		2学期	
A	国語↑	物理↑	国語↓	歴史↓
B	物理↑		国語↑	
C	生物↓	歴史↓		
D	算数↑		歴史↓	国語↑
E	地理↑			
F	歴史↓		国語↑	

可視化

- データを、人間が理解しやすいよう2次元（3次元）に配置する、あるいは画像や動画として表現すること



- 属性の軸を選択する、
項目同士の距離を調整する、等

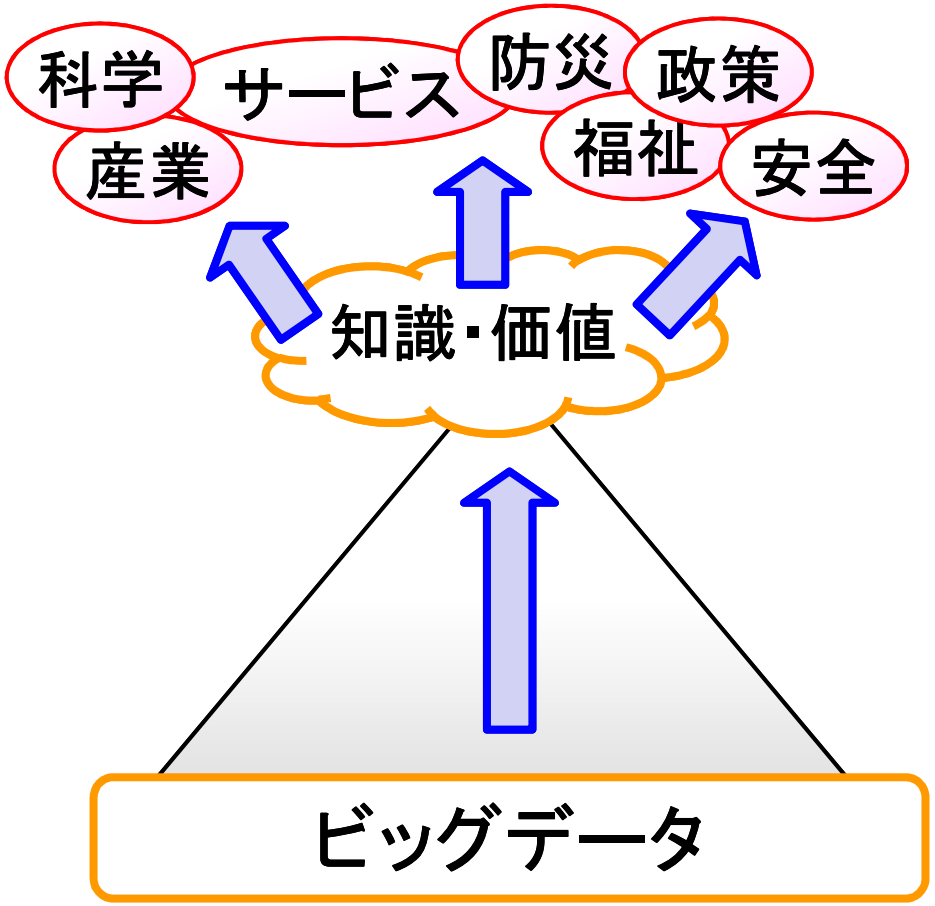
＋ グラフ描画
＋ 属性選択
＋ SOM
＋ 地図上への貼り付け
…

- 解析目的によって、見せ方が変わるはず

結局、何が大変なのか

- ビッグデータは難しくて、分かりにくいことが問題

- 人間の知識や認知と、
ビッグデータの間の
大きな乖離が原因

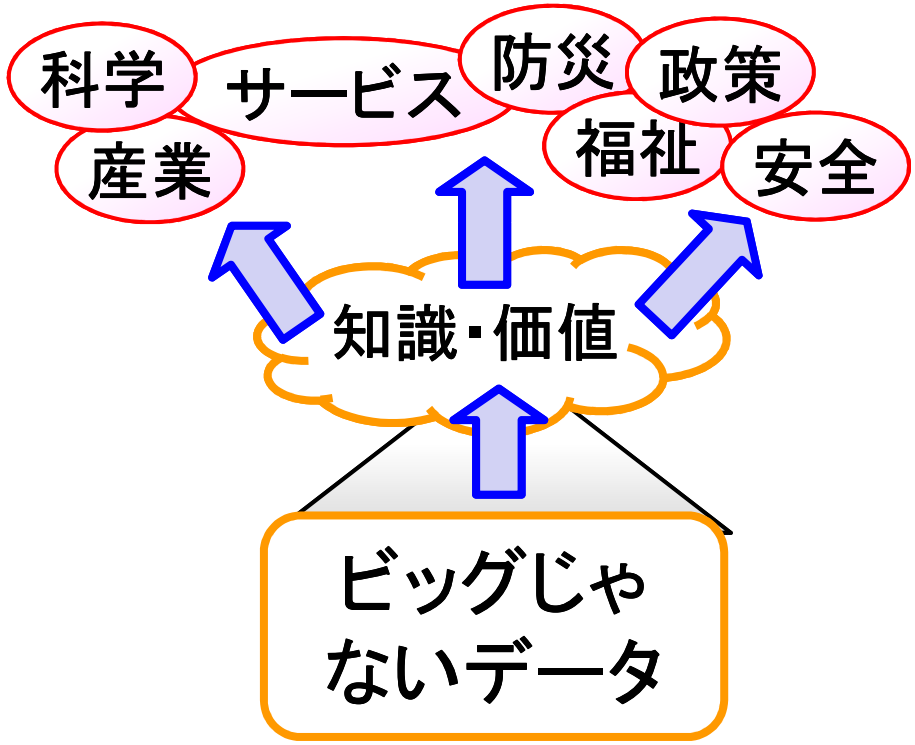


ビッグデータは何が大変か

- ビッグデータは難しくて、分かりにくいことが問題

- 人間の知識や認知と、
ビッグデータの間の
大きな乖離が原因

- 具体的、細かすぎる情報を
抽象化して、ビッグデータを
分かりやすく簡単にすれば、
全ては解決



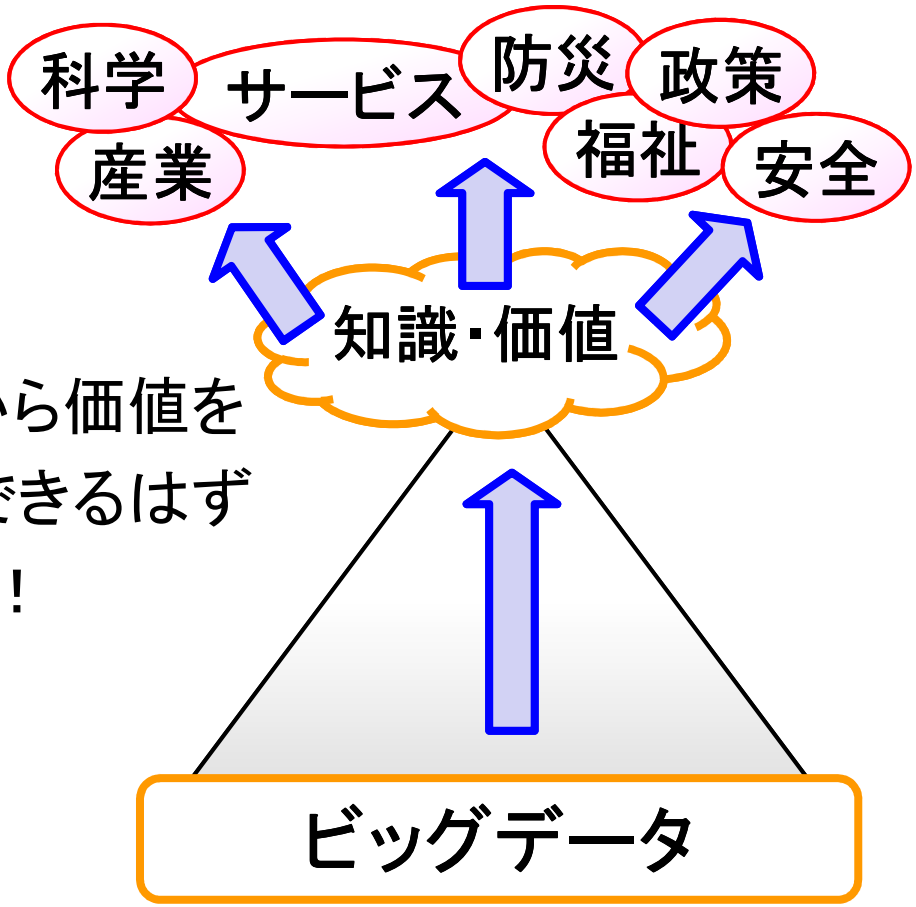
今、足りないもの

- 最終的には、価値を生み出したい
  ← 価値は**主観**によって定義される！

- 自分事として主観的に価値を
  創出する人が不足している

- 当事者なら、素人でも、データから価値を
  生み出す**「プロセス」**の設計はできるはず
  ➔ 1つのデータから100の価値！

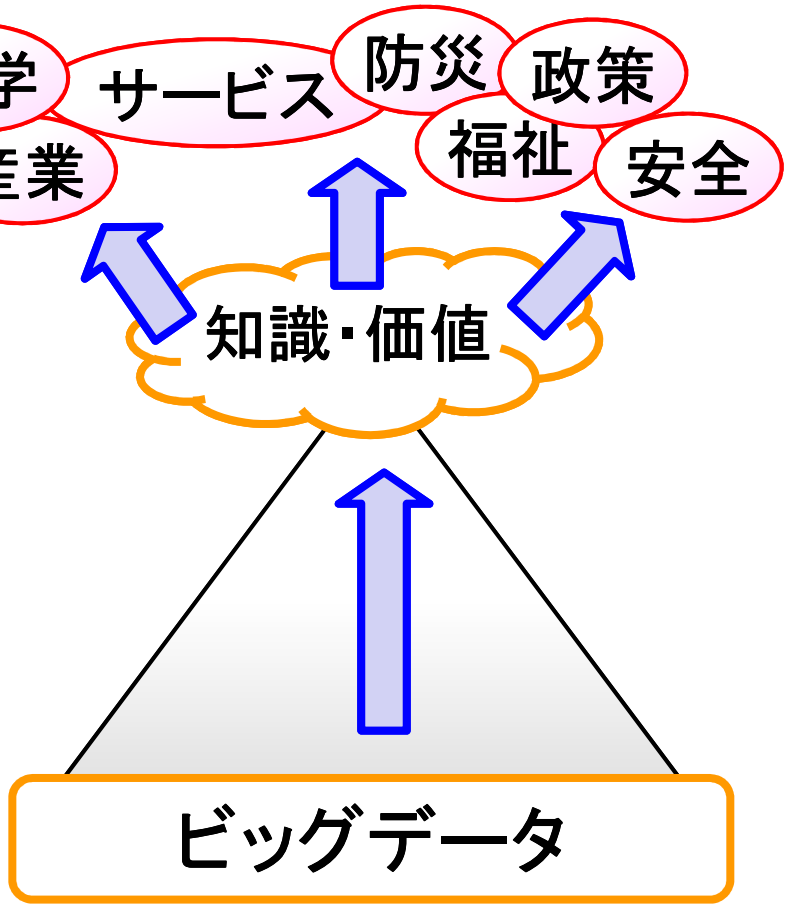
- データサイエンティスト不足が
原因とは、とても思えない



世界の中の日本

- 世界を見ても、1つのデータから100の価値が生まれているとは思えない　← やはり主観不足

- ただ、欧米は市場のサイズに大きな有利さがある
（資金：　日本500万、アメリカ5000万）

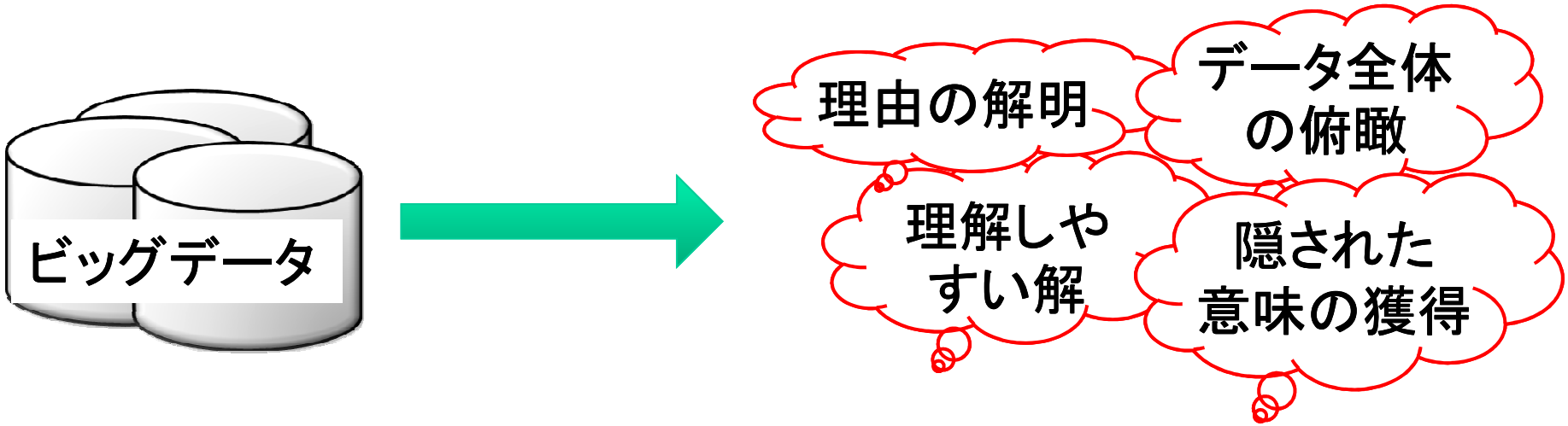


- ただし、主観を生む現場の強さにおいては、日本は世界トップレベル

- 本来は、世界を牛耳っていてもいいはず。それを助ける技術開発がとても重要

データ解析2.0

- データから物事と現象を深く理解できるようになる
- 箱物とブラックボックス技術に皮をかぶせる仕事から、
  データの意味を知り、価値を創造する仕事へ
- 小さな活動をする個人事業主やNPOもデータ解析
- 個人が自身の主観に基づく理解と解析結果を発信し、その集合知として社会が持つ「データに対する認識」が決まる時代



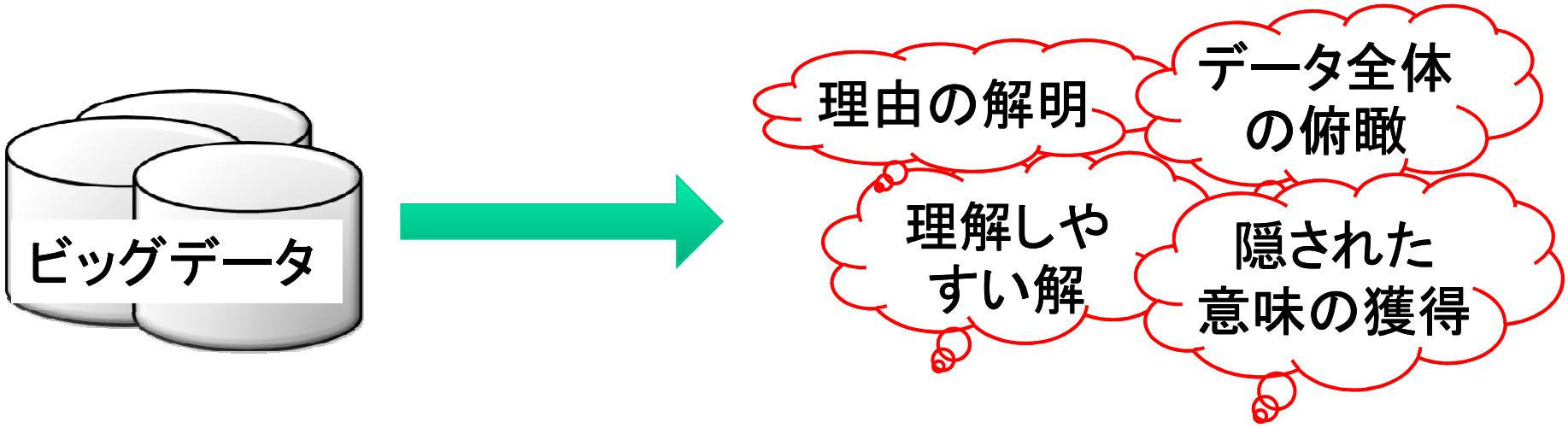
技術を知るためには

データ解析2.0の実現

- データ解析2.0時代を築くには、技術の利便性を高める必要

- そのためには、**「データ利用の定石」**が必要だが、
  そもそも目的もはっきりしないところで定石作りは無理
  （どの本を読めばいいですか？　→　そんなものはない！）

- そうなると、やはり技術に目が利く人が必要



技術の質の違いは、表面からは見えない

- 技術には、説明書きがある
  ＋ ●●ができます。××の精度がでます。実例は、、、

- 同じような目的に対して、異なる技術もある

- 外見上、同じようでも、結果の質が大きく異なることがある
  その違いは、**カタログスペックでは語りにくい**

- そこには**「技術の目利き」**が必要

結果：グラフカット

クラスタリング：データを類似・関連するグループに分ける問題

- グラフカットという方法は、分ける個数 **k** を指定
- 一般に広く使われているが、、、　新聞記事を分類すると、

＋ 見出しが「東京株、」で始まる記事が20以上に分割された
＋ ごちゃまぜのクラスタができている

小型機が着陸失敗、日本人女性重傷か…ネパール
東京株、午前終値は152円高 8カ月ぶり9700円台回復
オセロ松嶋が体調不良、笑福亭鶴瓶のみ出演へ
トヨタの前3月期、豊田社長ら首脳3氏が報酬1億円超
タケノコ採り4人無事 秋田・仙北で一時不明

結果：データ研磨

- 我が社の技術を使えばこの通り。グループ数の指定も不要

トヨタ14年3月期営業利益は1兆8000億円へ、市場予測下回る
ホンダの今期連結営業利益は前年比＋43.2%の見通し、市場予測下回る
NTTドコモの今期営業利益は0.3%増益を予想、市場予測をやや上回る
ファーストリテ、12年9―13年2月期営業利益は前年比＋5.3%
オリンパス、今期営業利益予想を前年比‐1.5%の350億円に下方修正
日産自、12年4―12月期営業益は18.4%減の3491億円
ソニー、今期当期損益予想を200億円の黒字で変更せず
ニコン、13年3月期営業利益予想を前年比‐40.1%に下方修正
三菱重が13年3月期営業益予想を上方修正、予測上回る

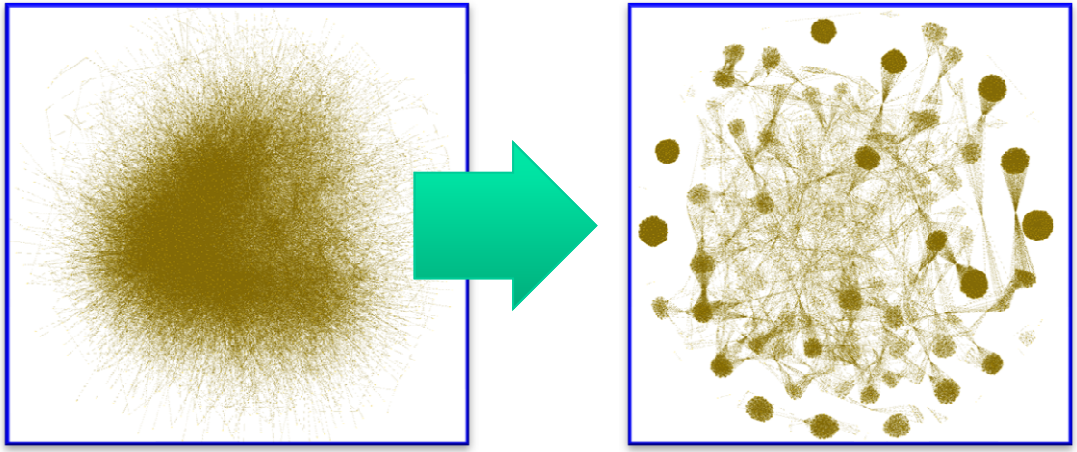
- 違いは、粒度とグループ数。グラフカットはグループ数が小さい問題が得意。我が社は細かく粒がそろっていると得意

データ研磨の威力

- 帝国データバンク様、企業間取引データ

  点：企業、線：取引ある、似てる：共通友達 PMI ≧0.6

	研磨前	研磨後
頂点数	3,282	3,282
枝数	35,168	73,132
クリーク数	32,953	343



買い物データを用いた、顧客が健康志向かどうかの予測精度

	クリーク	Newman	グラフカット
研磨前	60.60%	59.70%	60.03%
研磨後	**71.36%**	**62.76%**	**67.78%**

人工生成したベンチマークデータにおける、意味的構造の検出率

	研磨	Newman	グラフカット
ノイズ10%	**100.00%**	68.74%	76.10%
ノイズ40%	**99.69%**	7.91%	77.03%

違いを知る

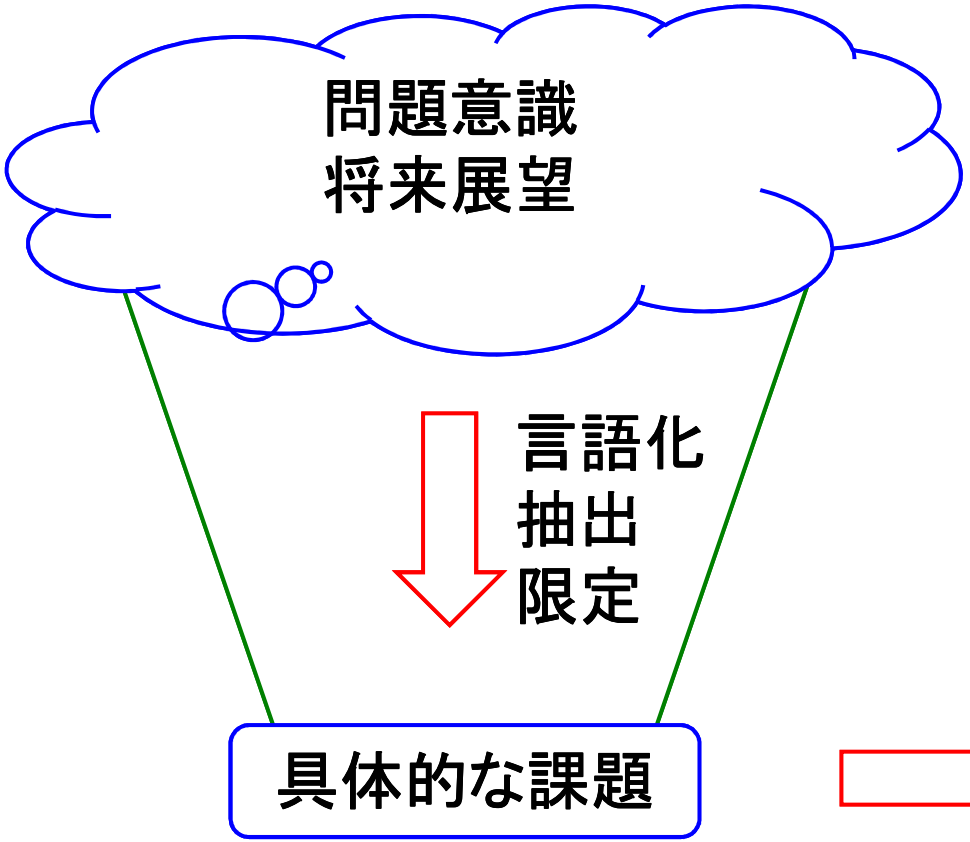
- このような違いは、素人にとって理解しがたく致命的
- かなり感覚的なことなので、文書化するのも難しい
  （つまり、教科書はない）

- 専門家に聞くしか、今のところは手がないように思える
  （将来的には、両方とも得意な何かが出てくるかも）

- そうなると、どこかからか知識を持っている人を連れてこないといけない
  （一番簡単なところは、同僚、友人
  しかし、専門家が一番良いだろう）

課題設定の難しさ

- ビジネスで一番大事なのは課題設定。そこを素人がしている



つまり「大変な課題」
を作っている

プロに依頼

発想自体にも落とし穴

- 音量センサーは防犯に使えるか
  → 音量で判別できるくらいの異常は、周りの人も気付く

- 騒音で迷惑する家の根拠に
  → 精度が悪い、主観で判断することを客観で判断

- 道を通った車の台数をカウントしよう
  → そこまでの精度はたぶん出ない

- 町の住民の民度や金持ち度合いを測ろう
  → 音量との相関は低いんじゃないかなあ

まとめ

- 「ビッグデータ流」の考え方、ものの見方
- 業務でデータ解析と向き合う、考え方
- 既存技術の理解のしかた、とらえ方
- 知識の手に入れ方、アイディアの出し方


ISP Networks
ISPは、人々の可能性に満ちた社会づくりに、技術と英知で貢献します
お問い合わせ：
info@isp-networks.com
http://www.isp-networks.com/