顧客分析・マーケティング・webログ解析などへの機械学習・統計的学習によるデータマイニングの基礎

2011年　SASユーザー総会
iAnalysis合同会社
最高解析責任者　倉橋一成

倉橋一成（Kurahashi　Issei）

- 博士課程：東京大学疫学・生物統計学教室（2011年3月卒）
- 現在：東大病院企画情報運営部
  - 最先端プロジェクト
  - 電子カルテのデータマイニング
- ブログ：Issei's Analysis ～おとうさんの解析日記～
  - http://d.hatena.ne.jp/isseing333/
- iAnalysis合同会社（アイアナリシス）
  - 最高解析責任者（CAO）
  - http://www.ianalysisllc.com/
  - 解析＆コンサル＆家庭教師しています
    - レセプトデータ
    - マーケティングデータ
    - 研究データ
    - 治験データ
    - 遺伝子データ

iAnalysis

Issei's Analysis ～おとうさんの解析日記～

- SASやRを使って統計解析について解説
- 「統計学を勉強するときに知っておきたい10ポイント」
  - http://d.hatena.ne.jp/isseing333/20110710/1310283922
  - 1,729ブックマーク！
  - 1日10,000ユーザーが訪問！！（ピーク時）
- Googleアナリティクスでアクセス状況のモニタリング
  - webページのアクセス解析を行うツール

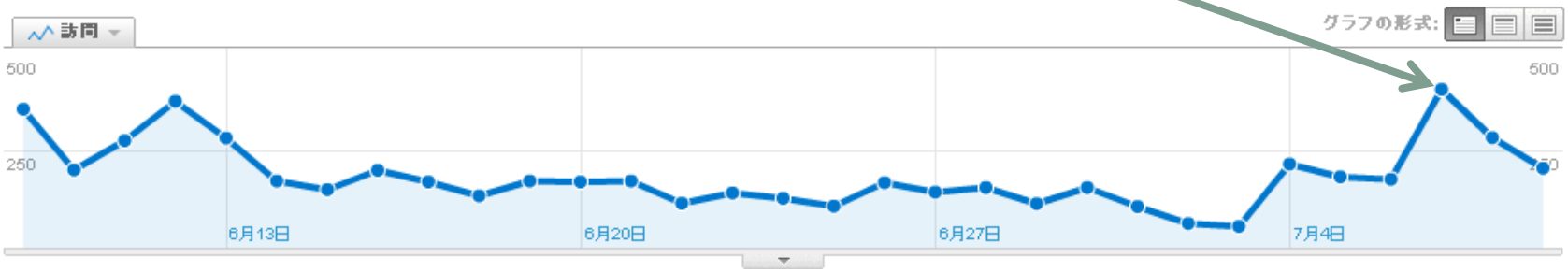
iAnalysis

Googleアナリティクスの推移

Before

マイレポート

最大アクセス数：410



Googleアナリティクスの推移

Before

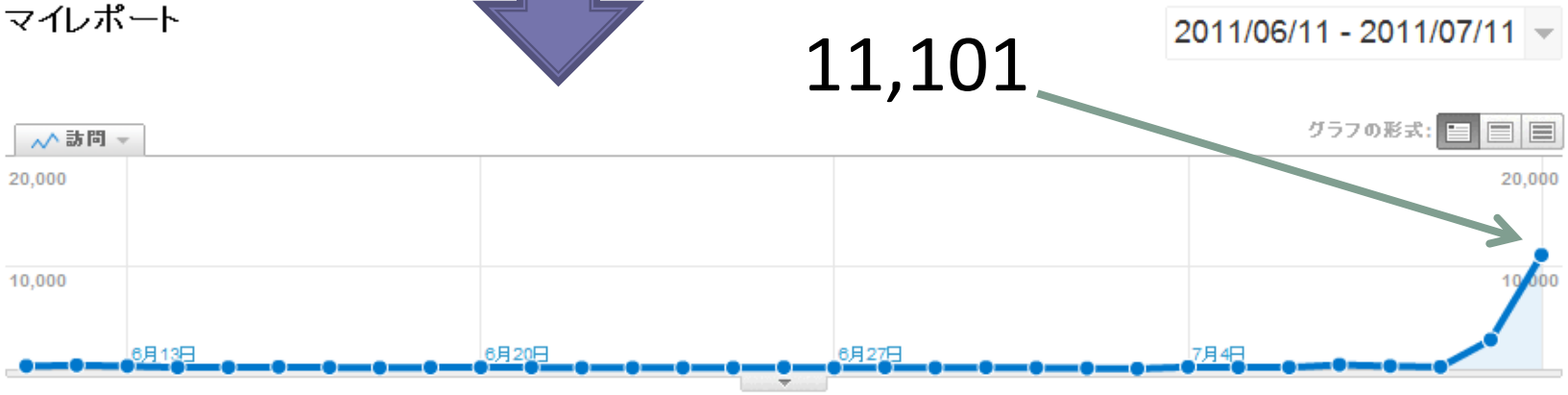
マイレポート

最大アクセス数：410

After(2日後)

マイレポート

11,101



iAnalysis

アクセス数の推移は時系列プロット

- 横軸：時間
- 縦軸：アクセス数

- データを「可視化」している
  - 可視化は「見える化」のこと
  - データを分析する上でとても重要

データを解析するための基本的な流れ

Phase I
データの収集・加工

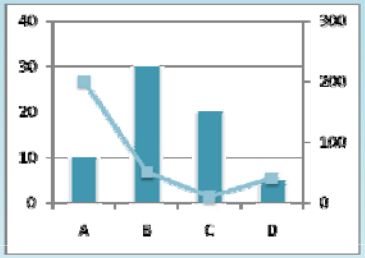
データベース

データはきちんと記録されていますか？

Phase II
データの可視化

散布図

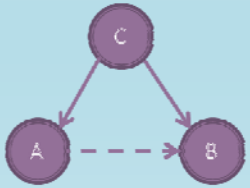
時系列プロット



棒グラフ

データを満足に分析できていますか？

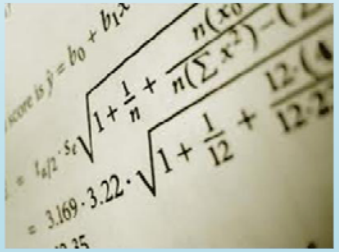
Phase III
データ解析



原因の分析

その効果は因果関係と言えますか？

Phase IV
効果測定デザイン



効果を見越したデザイン

介入による効果測定

効果測定のデザインが必要ではないですか？

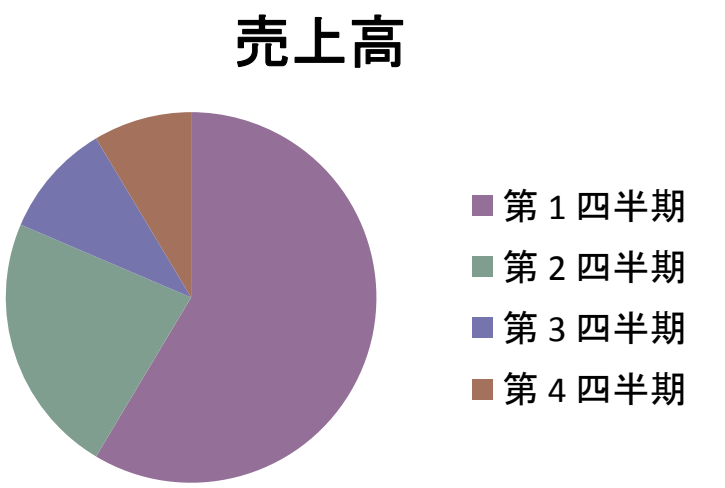
Phase I　データの加工・収集

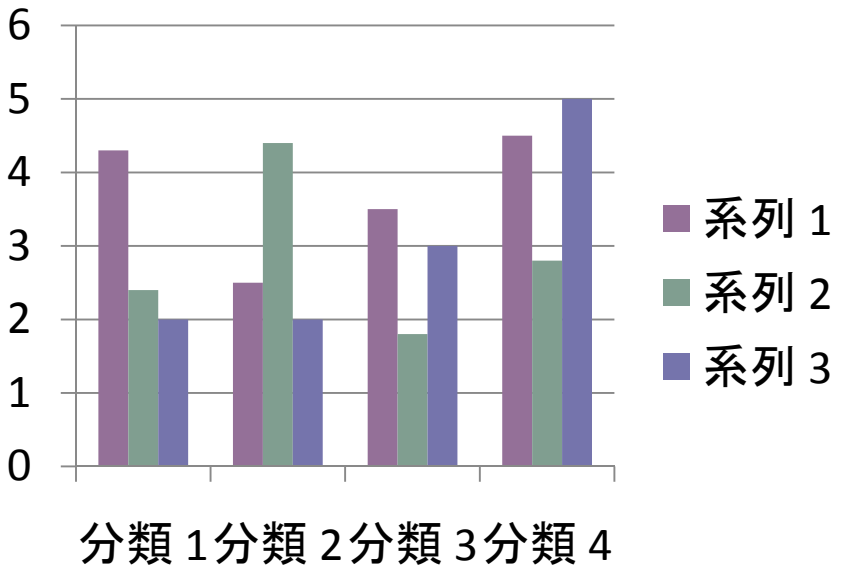
- データを「解析できる形」にする
  - データが全くない場合は作る
  - データが様々なデータベース（DB）に保存されている場合は統合する
    - DBに保存されていてもフォーマットが違う
    - 管理している部署が違う
    - アクセス制限がある
- “汚い”データを綺麗にクレンジングする必要がある
  - 欠測値の検討
    - そのまま残す、除去する、補完する
    - 補完：平均値、多重補完（proc MI）
  - 外れ値の検討
    - 間違った値→除去、欠測
    - 間違ってないが解析には大きすぎる→ウィンザライゼーション

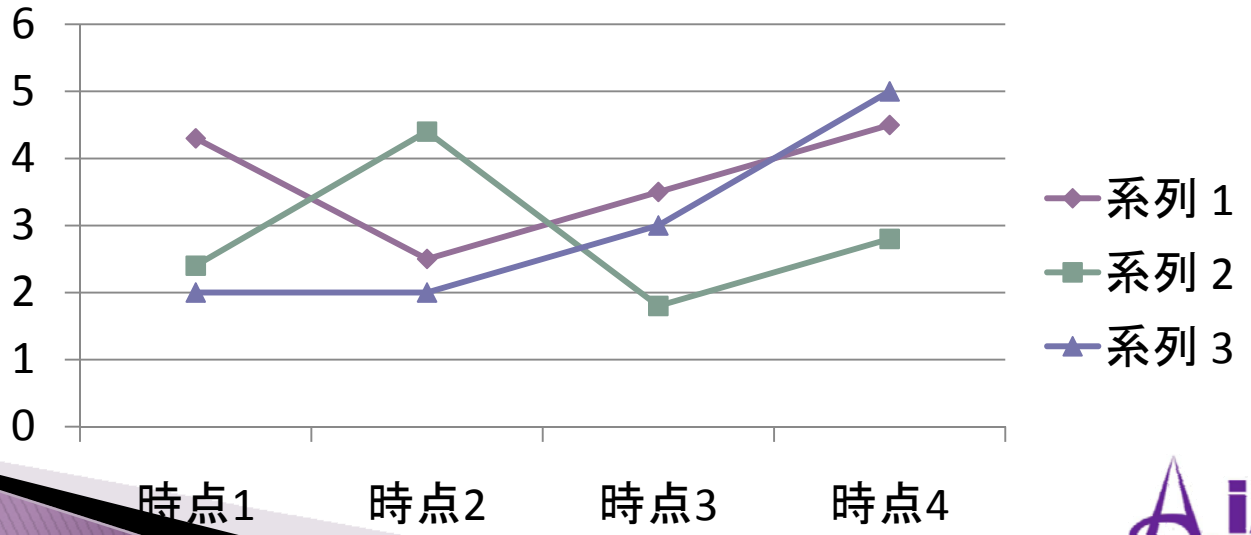
Phase II　データの可視化

- 数値の羅列であるデータを目に見える形にする
  - ヒストグラム
  - 箱ひげ図
  - 棒グラフ
  - 円グラフ
  - 時系列プロット
  - 散布図
- 「データ」とは変数（列）×レコード（行）の事
  - 変数の例：性別、年齢、課金、収入、アクセス数、時間
  - レコードの例：人、チェーン店の店舗

可視化の例


売上高
第 1 四半期
第 2 四半期
第 3 四半期
第 4 四半期


6
5
4
3
2
1
0
分類 1
分類 2
分類 3
分類 4
系列 1
系列 2
系列 3


6
5
4
3
2
1
0
時点1
時点2
時点3
時点4
系列 1
系列 2
系列 3

Gapminder

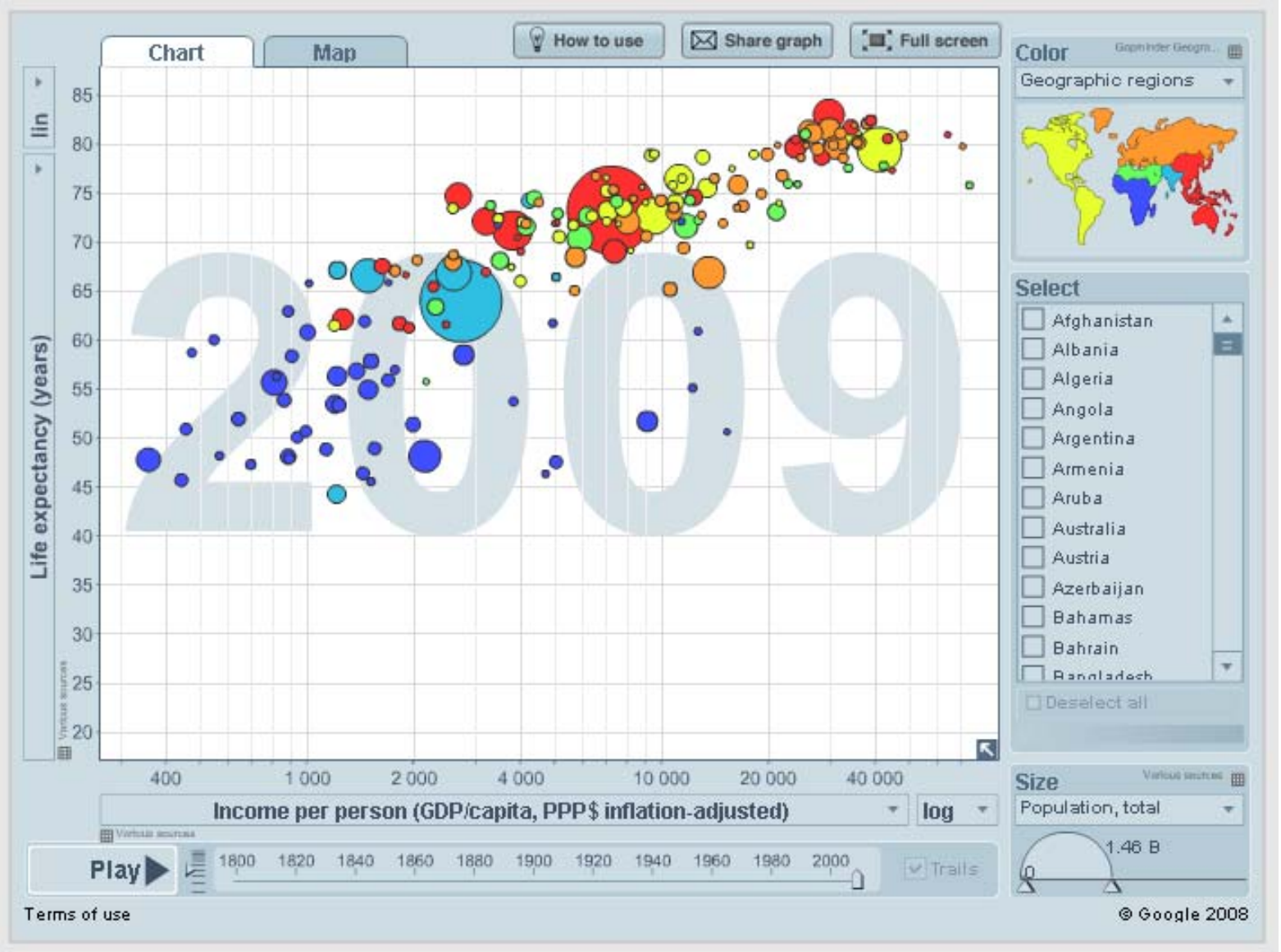


Chart
Map
How to use
Share graph
Full screen
Color
Geographic regions
Select
Afghanistan
Albania
Algeria
Angola
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Deselect all
Size
Population, total
1.46 B
lin
Life expectancy (years)
85
80
75
70
65
60
55
50
45
40
35
30
25
20
2009
400
1 000
2 000
4 000
10 000
20 000
40 000
Income per person (GDP/capita, PPP$ inflation-adjusted)
log
Play
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Trails
Terms of use
© Google 2008

Phase III　データ解析

- データに様々な「モデル」を当てはめて、情報を探索する
  - 変数同士の関連をチェックする
  - ある変数に影響を与えている変数は何か？
    - マーケティング：購買、課金、リピートに影響する変数は何か？
    - 医療：疾病発症、生存時間、再発、予後に影響する変数は何か？
- 予測したい変数：結果変数
- 予測に使われる変数：説明変数
  - 呼び方がたくさんある
    - 結果変数→応答変数、従属変数
    - 説明変数→予測変数、独立変数
- モデルの例：「購買したかどうか」という結果変数を、「性別・年齢・居住地域・収入」などの説明変数で予測する
  - ロジスティック回帰、判別分析、SVM、ニューラルネットワーク

iAnalysis

様々なモデル

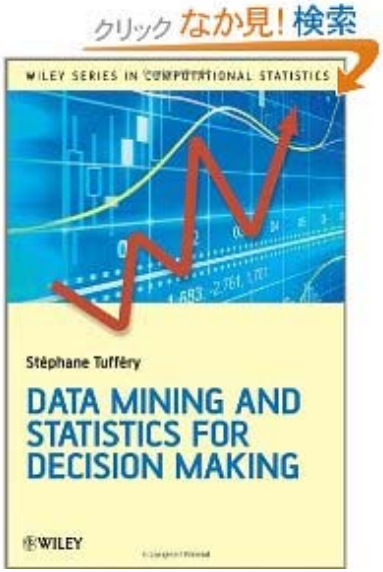
- どういう人が商品を購入するか？
- logit(購入する確率) = a × 性別 + b × 年齢 + …
  - ロジスティック回帰
  - logit：対数オッズ
  - 購入する確率をモデル化している
- どういう人がいくら課金するか？
- 課金の額 = a × 性別 + b × 年齢 + …
  - 線形回帰
  - 消費した金額をモデル化している
- どういう人が何回購入したか？
- log(購入回数) = a × 性別 + b × 年齢 + …
  - ポアソン回帰
  - 購入した回数をモデル化している

iAnalysis

Phase IV　効果測定デザイン

- 既にあるデータを分析するだけでは「介入効果」は測定不可
  - 広告の購買効果
  - 薬剤の治療効果
- ランダム化試験を行う
  - 対象者をランダムにいくつかの群に分けて、異なる介入を行う
  - 介入後の結果を比較する
- 例
  - 広告A vs. 広告B
  - 薬剤A vs. 薬剤B
- サンプルサイズ設計を行い、介入を行うべき人数を計算する
  - ランダム化試験を行えば、介入人数は数百人で十分な場合が多い
  - 大がかりな試験をすることなく、介入効果の有無を見積もることが可能

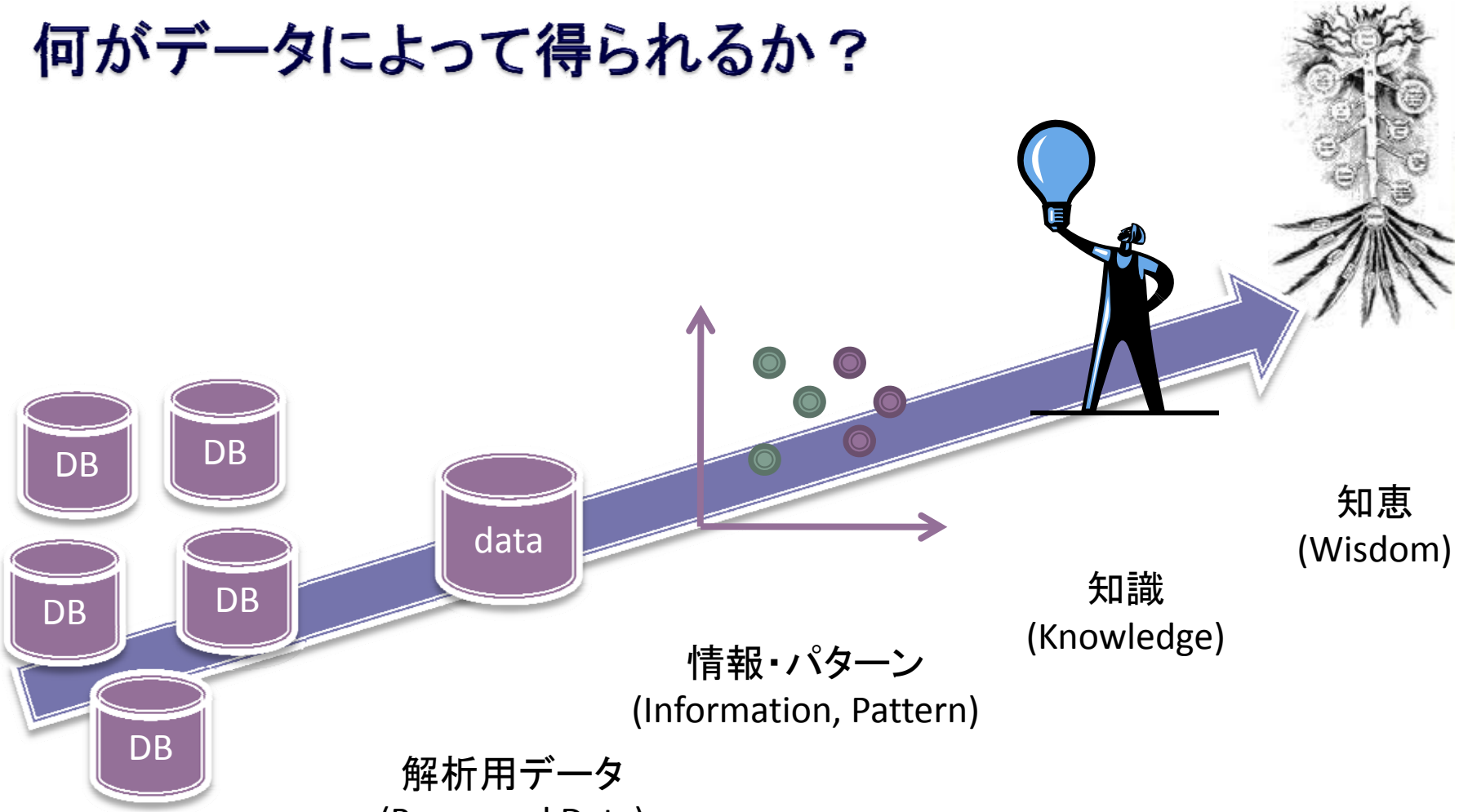
本題:データマイニングとは？



“Data Mining and Statistics for Decision Making.”
『意思決定のためのデータマイニングと統計学』
Stphane Tuffry

Data mining is a tool for extracting the jewel of truth from the data.
データマイニングとはデータから真実という宝を抽出するためのツールである

何がデータによって得られるか？



システムエンジニア (SE)

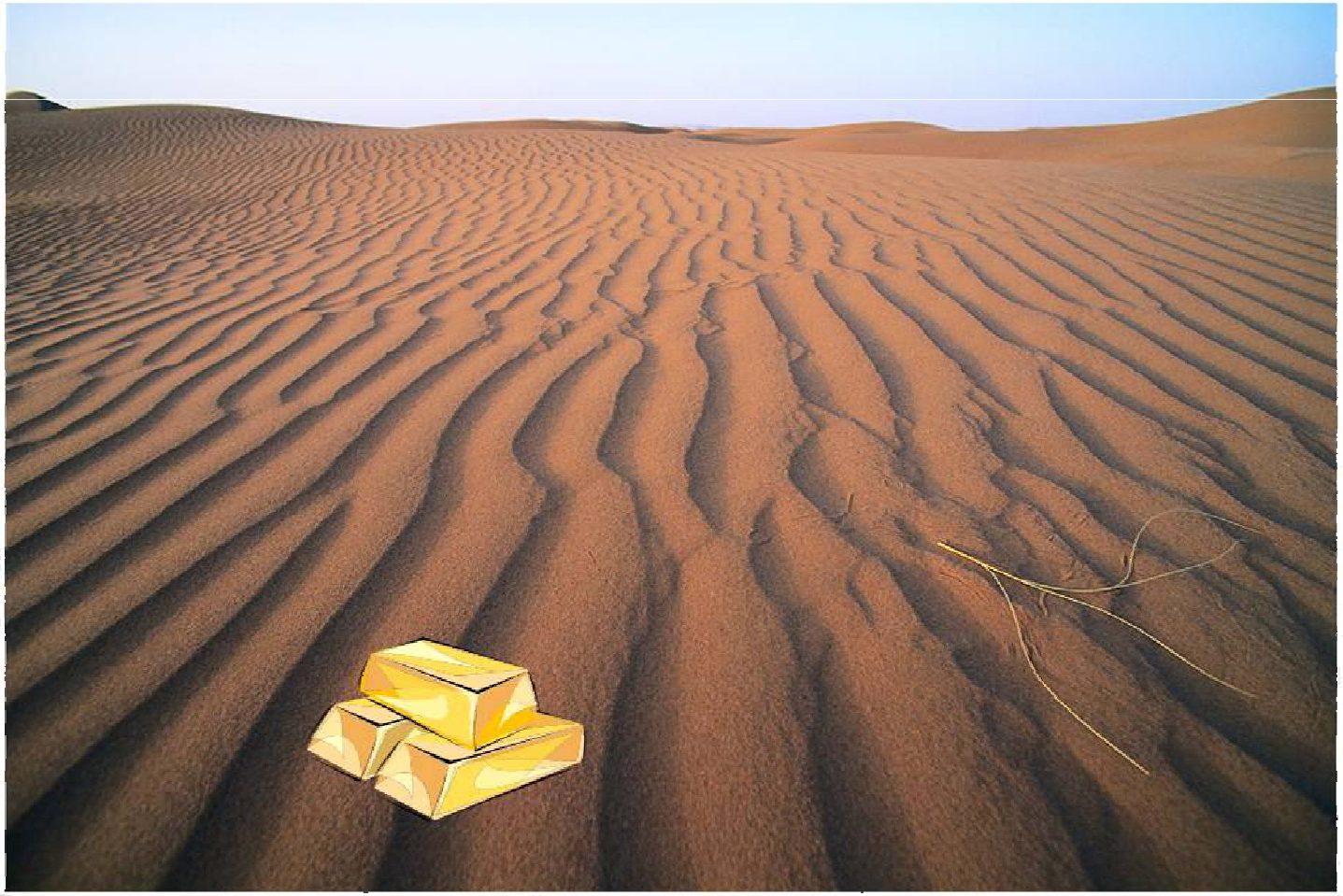
コンサルタント (Consultant)

統計家 (Statistician)

クライアント (Client)

どうやって大量のデータから金を見つけますか？

How do you find the gold?



http://www.ianalysisllc.com/

iAnalysis

活用例1. CRM (customer relationship management)

- 顧客関係管理
  - 顧客との関係を長期的に改善、維持することにより企業の収益を最大化するという経営戦略
  - http://www.atmarkit.co.jp/fitbiz/serial/datamining/01/01.html
- 顧客の情報(データ)を分析することで関係を強化し、利益につなげる
- 顧客の特性と分析手法
  - コスト:新規顧客＞顧客維持
    - 長期間顧客である人の特性を探る
    - クラスター分析、コレスポンデンス解析
  - コスト:離反顧客の引き戻し＞離反の防止
    - 離反する、しないを予測する
    - ロジスティック回帰、生存時間解析
  - 一部の顧客が大量消費することがある
    - そのような消費者の把握、囲い込み
    - データの可視化

iAnalysis

活用例2. RFM (recency, frequency, monetary)

- 顧客の購買行動・購買履歴から、優良顧客のセグメンテーションなどを行う顧客分析手法
  - R：最新購買日、いつ購入したか、最近購入したか
  - F：累計購買回数：どのくらいの頻度で購入したか
  - M：累計購買金額：いくら消費しているか
- RFMそれぞれセグメント化
  - 集計、スコアリング
- 対象者のデモグラフィック（背景）からスコアを予測
  - Web調査
  - チェックインのデータ

活用例3. 迷惑メールのフィルタリング

- Gmail
  - 「迷惑メール」であるかどうか予測し、フィルタを行う
    - ロジスティック回帰
    - ナイーブベイズ
  - 利用者が「迷惑メール」と選択した時点で予測方法を更新する
    - ベイズによるパラメータ更新
- Google:「次の10年で熱い職業は統計学」
  - あらゆるデータが記録される時代
  - データをどのように有効活用するか！
  - http://www.publickey1.jp/blog/10/10_3.html

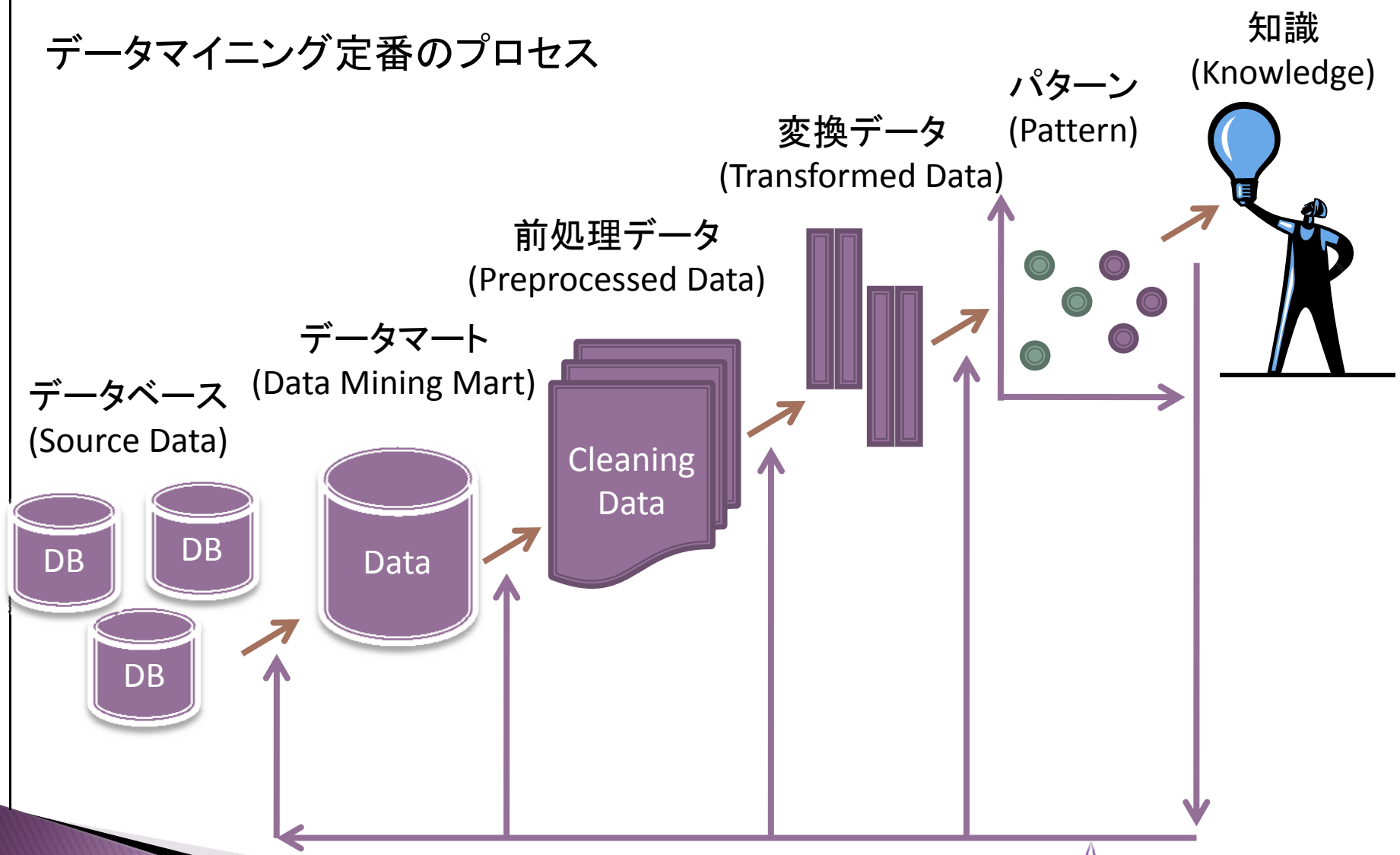
活用例4. リコメンド (recomend)

- 電子商店などで、ユーザの好みを分析し、各ユーザごとに興味のありそうな情報を選択して表示するサービスのこと
- Amazon
  - Webサイトで顧客層ごとに異なるトップメニューを用意
  - ある商品を購入したら他の商品を推奨する
- マクドナルド
  - 会員の購買履歴を分析して個々人に異なるキャンペーンを行う
  - おさいふケータイを利用している1,000万人が対象
  - 7/14 日経新聞朝刊
    - 日経web版の記事が消えてる？

他にも様々な分野で利用されている

- セイバーメトリクス
  - 野球データを分析
- ワインの品質
  - アッシェンフェルターの方程式
  - 12＋0.00117×冬の降雨量＋0.0614×育成期平均気温＋0.00386×収穫期降雨量
- Google物価指数
  - 世界中のショッピングデータから独自に景気動向指数を計算

Knowledge Discovery in Data (KDD) Process



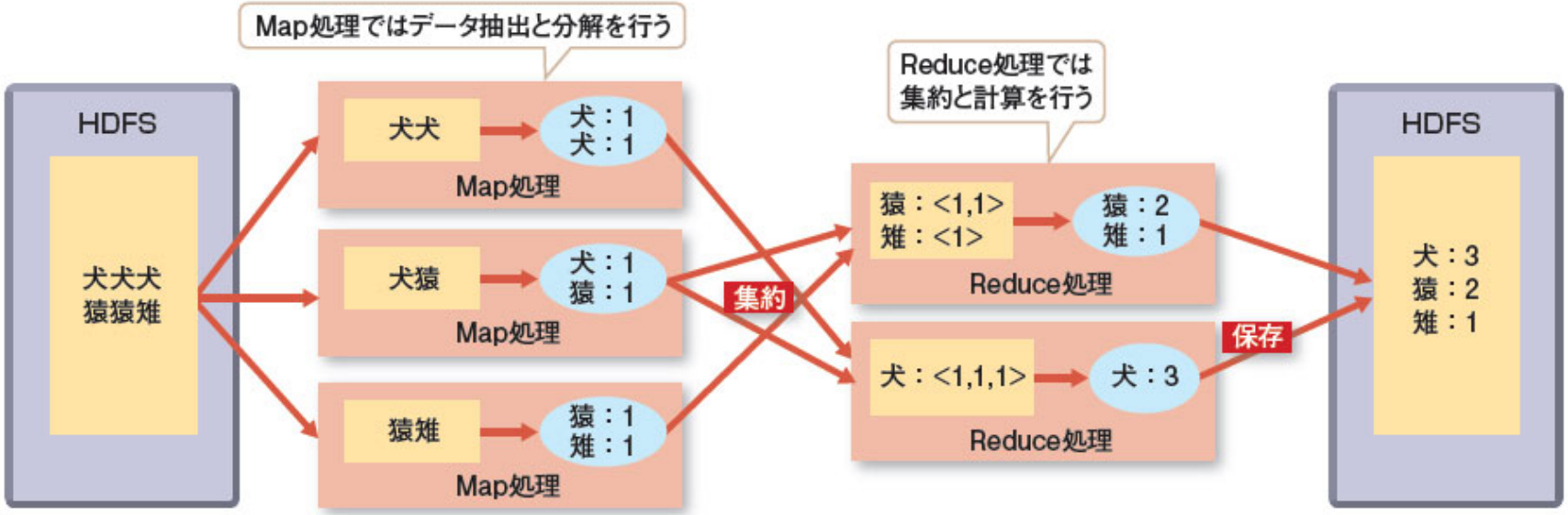
データベース→データマート

- 様々なところに記録されているデータを統合する
  - 社内の部署連携
  - データベースエンジニア
- 大規模データを扱う必要
  - 数100GB～数10TB
  - Facebookは1日に約100TBのデータが発生
  - Googleは約200億（？）のサイトから検索を行っている（約400TB？）
  - Amazonは数千万アイテムの中からリコメンド（推奨）している
- 「分散処理」によって高速に処理を行う
  - Hadoop（ハドゥープ）
    - Googleの基盤技術であるMapReduceをJavaでオープンソース実装した分散処理のフレームワーク

iAnalysis

複数のコンピュータに処理を分散させて並列計算

- Hadoop
  - 分散処理のためのJavaライブラリ
- 「Hadoop分散処理、6時間から5分に高速化 - Yahoo! Japan」
  - http://journal.mycom.co.jp/news/2009/03/04/030/index.html
- MapReduceという概念
  - Map：データを分散して処理
  - Reduce：処理を統合



http://itpro.nikkeibp.co.jp/article/COLUMN/20110112/355999/

集計はもちろん高度な統計解析も分散処理できる！

- 論文：Map-Reduce for Machine Learning on Multicore
  - http://www.cs.stanford.edu/people/ang//papers/nips06-mapreducemulticore.pdf
- MapReduceできる統計手法
  - 局所重み付き線形回帰（Locally Weighted Linear Regression）
  - ナイーブベイズ（Naive Bayes）
  - 判別分析（Gaussian Discriminative Analysis）
  - K-means法
  - ロジスティック回帰（Logistic Regression）
  - ニューラルネットワーク（Neural Network）
  - 主成分分析（Principal Components Analysis）
  - 独立成分分析（Independent Component Analysis）
  - EMアルゴリズム（Expectation Maximization）
  - サポートベクターマシン（Support Vector Machine）
- これらを組み込んでいるのがMahout

簡単な線形回帰を分散処理

- ダミーデータ
  - x：100個の乱数
  - y：2x + 誤差
- データを分割する
  - x_sub： 1～50、51～100
  - y_sub： 1～50、51～100
  - xy平方和_sub： sum(y_sub * x_sub)
  - xx平方和_sub ：sum(x_sub * x_sub)
- 統合する
  - (xy平方和_sub1 + xy平方和_sub2)/(xx平方和_sub1 + xx平方和_sub2)
- 全体でパラメータの推定と一致する
  - yx平方和 / xx平方和

分散処理の応用可能性

- 大規模データでの解析
- 分散しているデータの統合
  - データ自体は個人情報の問題で持ち出せない
  - 集計したデータは外に出せる
  - 後で統合できる状態まで計算して、集計情報を提供する
  - データ全体を統合して計算したものと同じ結果が得られる

データマイニングの話に戻ります

- データの前処理、クレンジング
  - データを解析できる形に加工する
    - サンプリング
    - 欠測値の処理
    - 外れ値の処理
    - カテゴリ化
    - ダミー変数の作成
- 変数変換
  - 回帰係数が同じスケールになるような変換
    - 最大・最小標準化、Zスコア
  - なるべく正規分布に近くなるような変換、分散安定化変換
    - 対数変換、平方根変換、逆正弦変換
  - 非線形の影響を確認したい
    - 2乗、3乗、スプライン、対数、平方根

大枠でのデータマイニングの分類

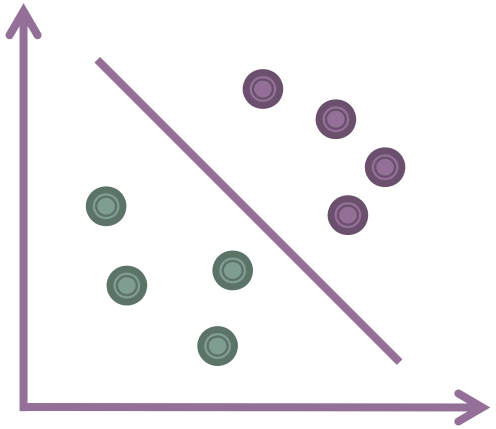
- 予測的データマイニング、教師付き学習
  - 過去のデータを使って将来の値を予測する
    1. 回帰モデル
    2. クラスタ予測
    3. 機械学習
- 記述的データマイニング、非教師付き学習
  - データのパターンを発見する
    A) 関連ルール（association rule）
    B) クラスタリング
    C) テキストマイニング

1. 回帰モデル

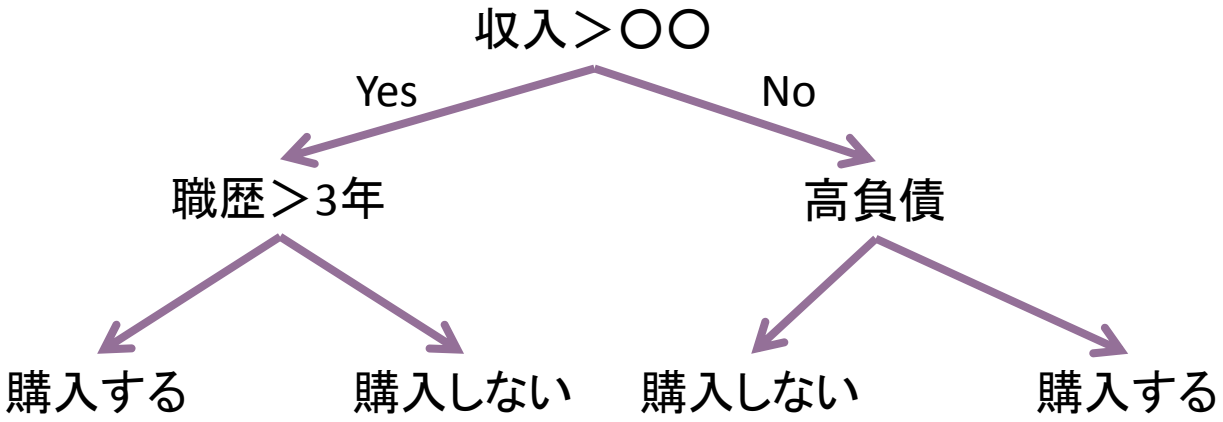
- 一般化線形モデル
  - 線形回帰：$Y = a + bx_1 + cx_2 + \ldots$
  - ロジスティック回帰：$\mathrm{logit}(P) = a + bx_1 + cx_2 + \ldots$
  - ポアソン回帰：$\log(Y) = a + bx_1 + cx_2 + \ldots$
- 正則化回帰
  - リッジ回帰、LASSO、LARS
    - パラメータ推定にペナルティを付けて過適合を防ぐ
- 時系列モデル
  - ARIMA
- 非線形回帰
  - スプライン、一般化加法モデル
  - 多変量加法回帰スプライン（multiple adaptive regression splines; MARS）
- 生存時間解析
  - カップラン・マイヤー曲線、Cox比例ハザードモデル

2. クラスタ予測

- 判別分析
- ロジスティック回帰

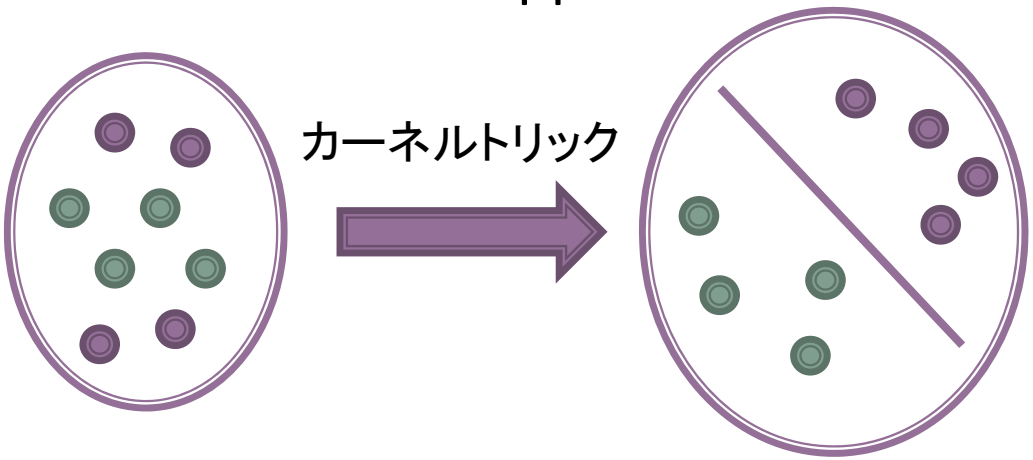


- 決定木
  - 再帰的分割アルゴリズム（Recursive Partitioning Algorithms）

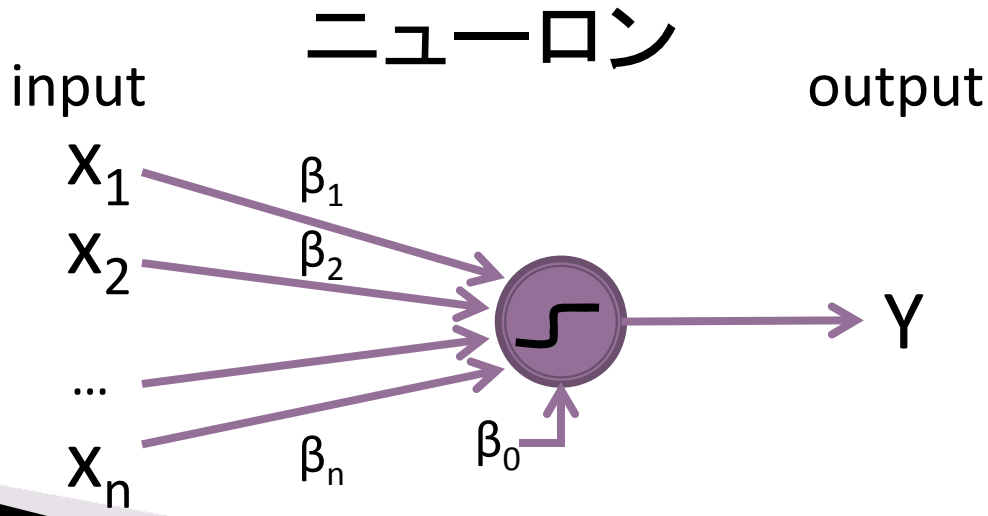


3. 機械学習

- サポートベクターマシン（Support Vector Machine; SVM）



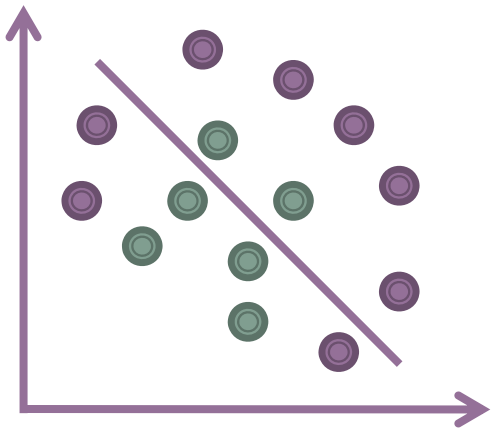
- ニューラルネットワーク（Neural Network; NN）



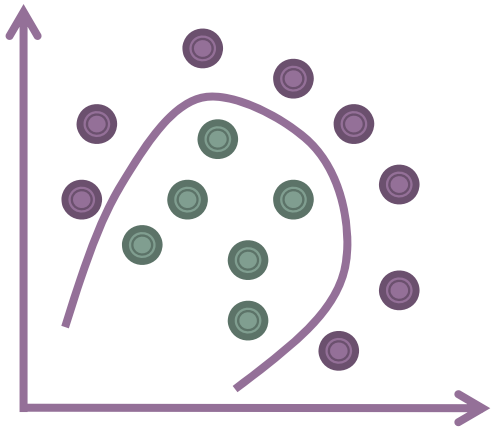
iAnalysis

機械学習の利点：SVMとNNでは非線形な予測が可能

- 線形予測・判別



- 非線形予測・判別

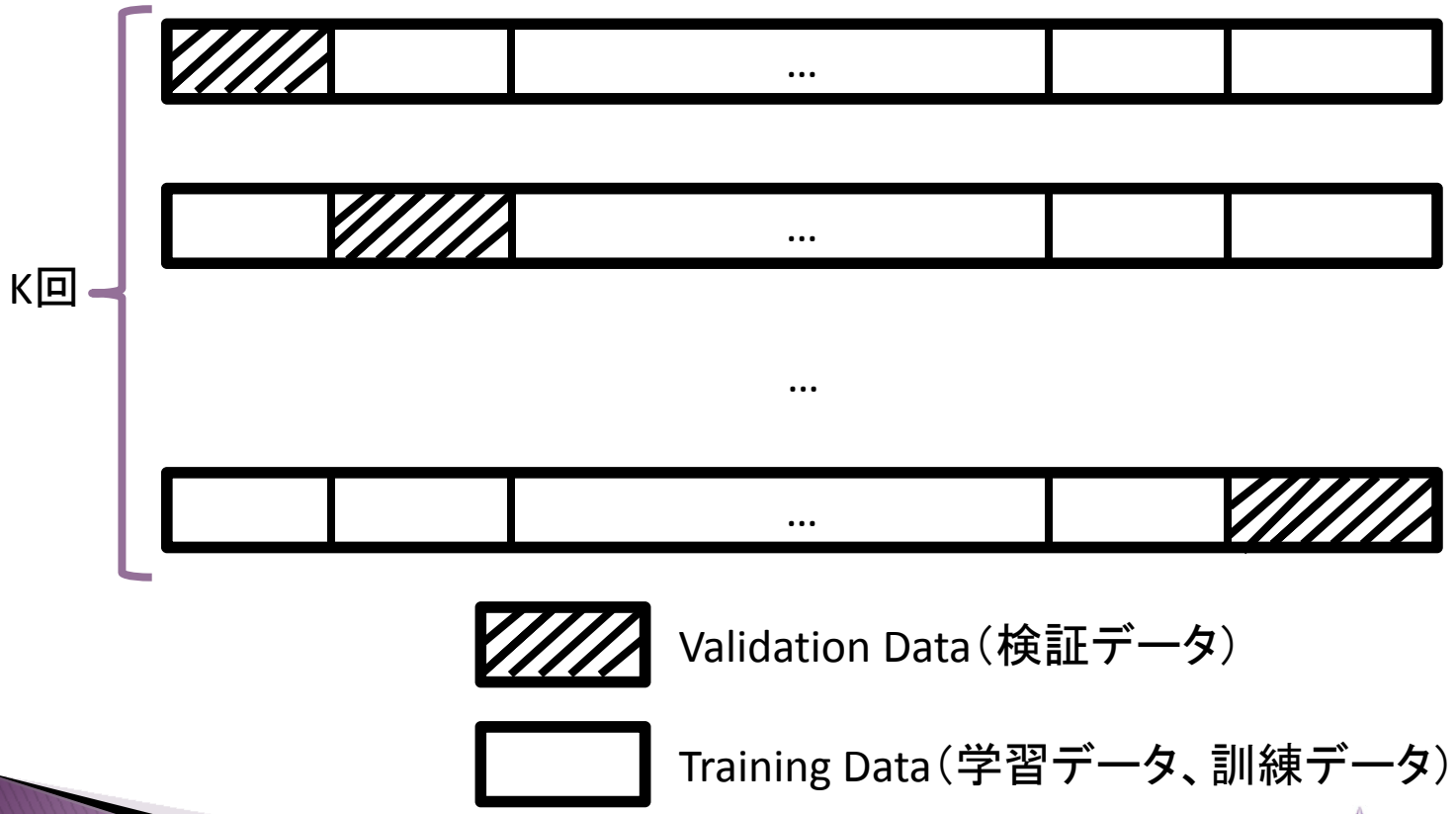


機械学習の欠点

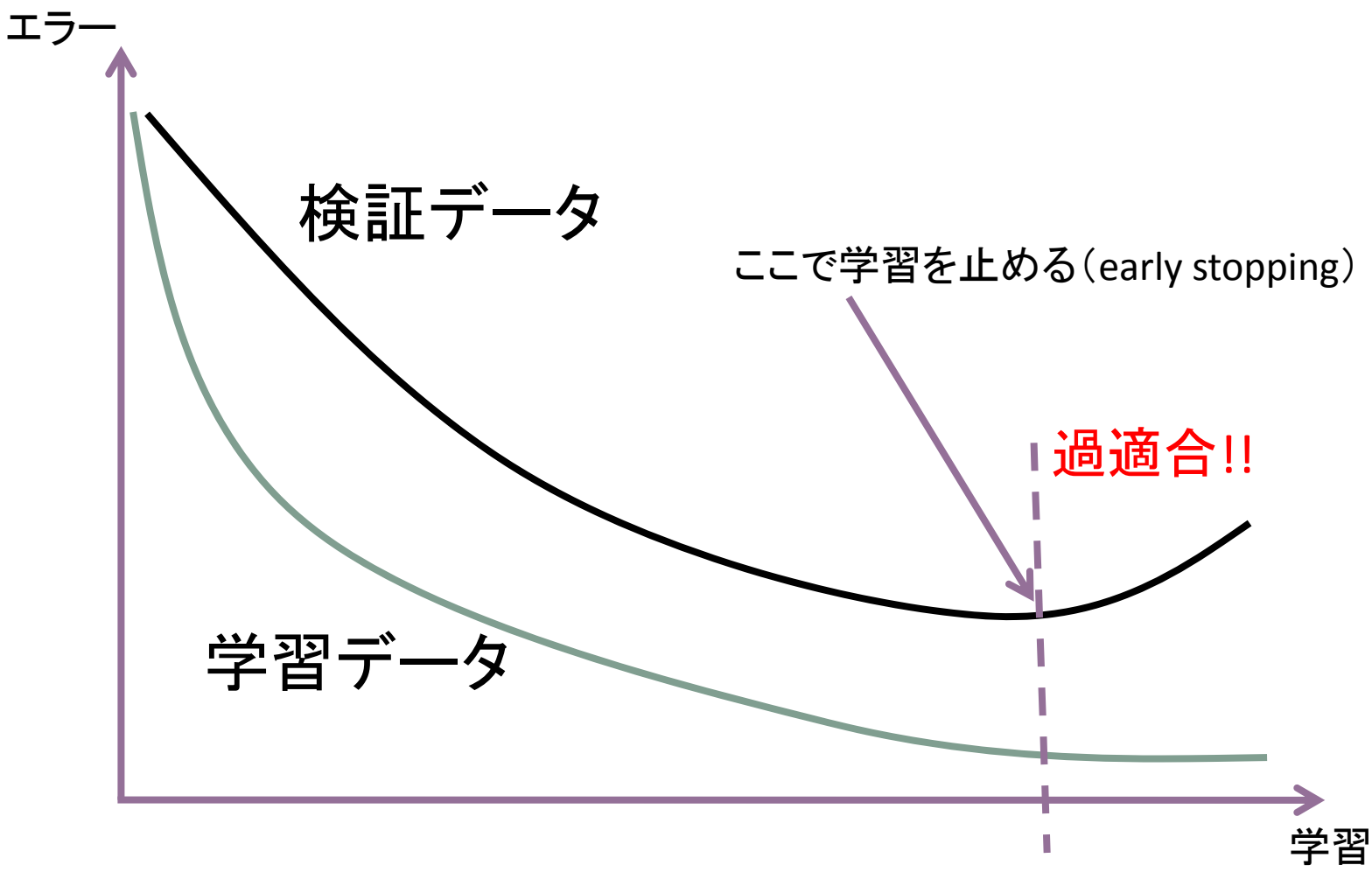
- 複雑な結果（非線形）になるので解釈が難しい
- データの特徴を単純化（モデル化）できない
- 複雑なモデルから解釈可能なルールを抽出する方法もある
  - 決定木を駆使してルールを抽出する

教師付き学習で気を付ける事：過適合（overfitting）

- 「学習」させ過ぎると「過適合」が起こる
- クロスバリデーション（交差検証）を行う必要がある
  - K-fold Cross-Validation（K=10、K=2、K=nの場合が多い）



学習と過適合の関係



A. 関連ルール

- バスケット分析
  - 一緒に購入しやすいアイテムのパターンを抽出する

・support(X→Y) = number of (X U Y) / total number
・confidence(X→Y) = support(X U Y) / support(X)

対象者	アイテム
1	Item6, Item2, Item4
2	Item1, Item3
3	Item3, Item6
4	Item1, Item2, Item3
5	Item1, Item2, Item6, Item3, Item4
6	Item2, Item6, Item4
7	Item2, Item4, Item6

Item6 → Item4, Item2　　support = 4/7, confidence = 4/5
Item6, Item2 → Item4　　support = 4/7, confidence = 4/4

iAnalysis

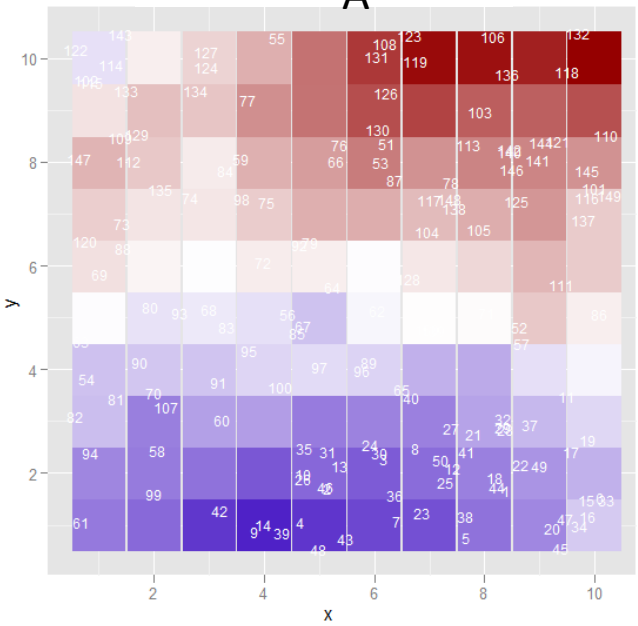
B. クラスタリング

- K-means
  - ↓このサイトが非常に分かりやすい
  - http://d.hatena.ne.jp/nitoyon/20090409/kmeans_visualise
- 自己組織化マップ（Self-Organization Map; SOM）
  - ニューロンを使ったクラスタリング
- ネットワーク分析
  - データのネットワーク図を描いて特徴を探る

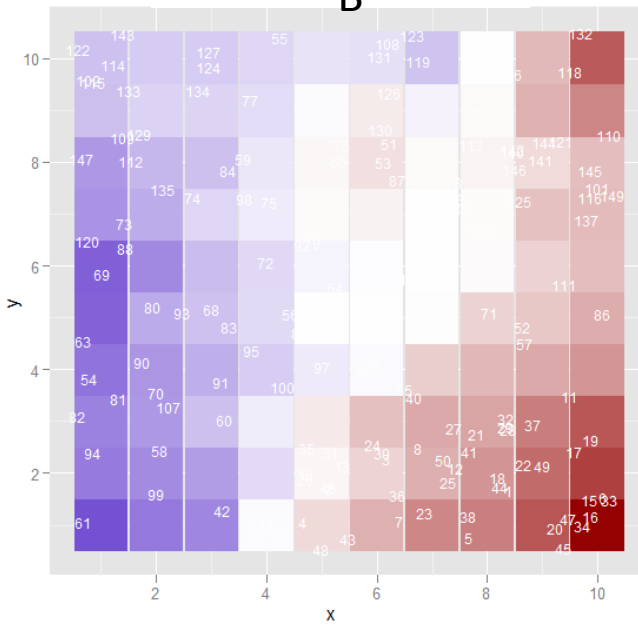
自己組織化マップ

http://d.hatena.ne.jp/isseing333/20110224/1298512568

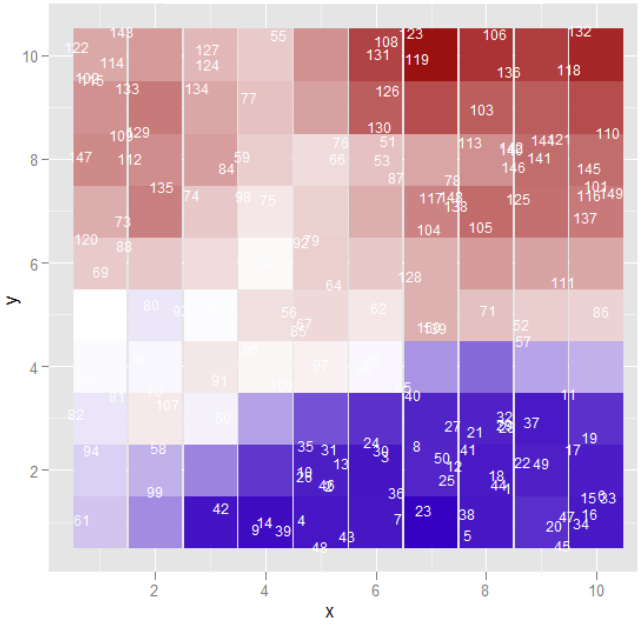
A



B



C



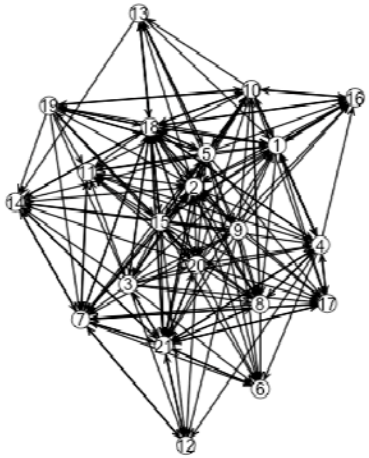
AとCの関連は強い
AとBの関連は弱い

iAnalysis

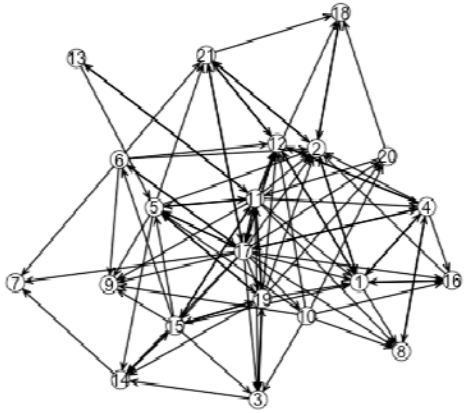
ネットワーク分析

- ハイテク企業の管理職21人の社会ネットワーク
  - 『ネットワーク分析 (Rで学ぶデータサイエンス 8) 』

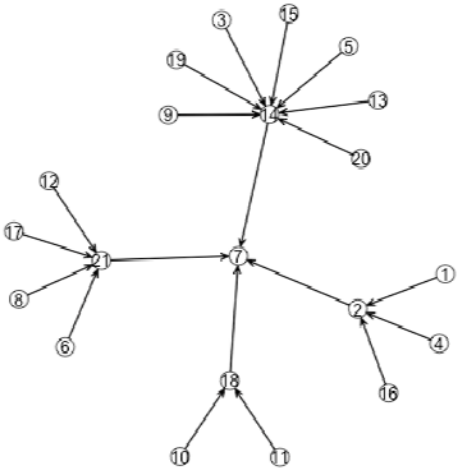
「アドバイスを求める」



「友人である」



「報告をする」



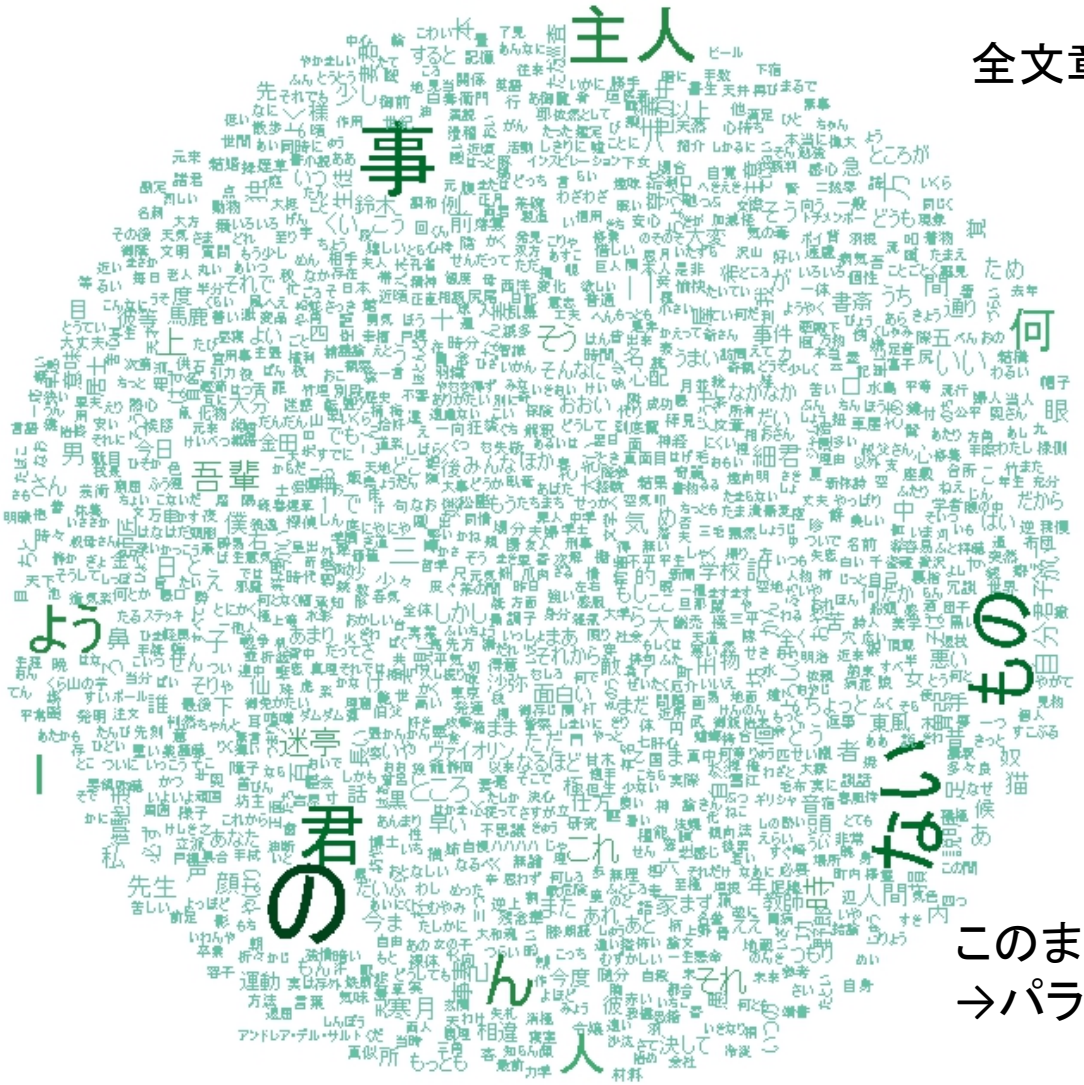
http://d.hatena.ne.jp/yokkuns/20110223/1298416018

iAnalysis

C. テキストマイニング

- 「文章」データから情報を抽出する
  - Twitter、Facebook
  - 小説
- ワードクラウド
  - 形態素解析＋単語が利用されている頻度の可視化

例:「吾輩は猫である」のワードクラウド



全文章(約35万文字)を分析

このままではちょっと分からない
→パラメータを変えたりして「マイニング」

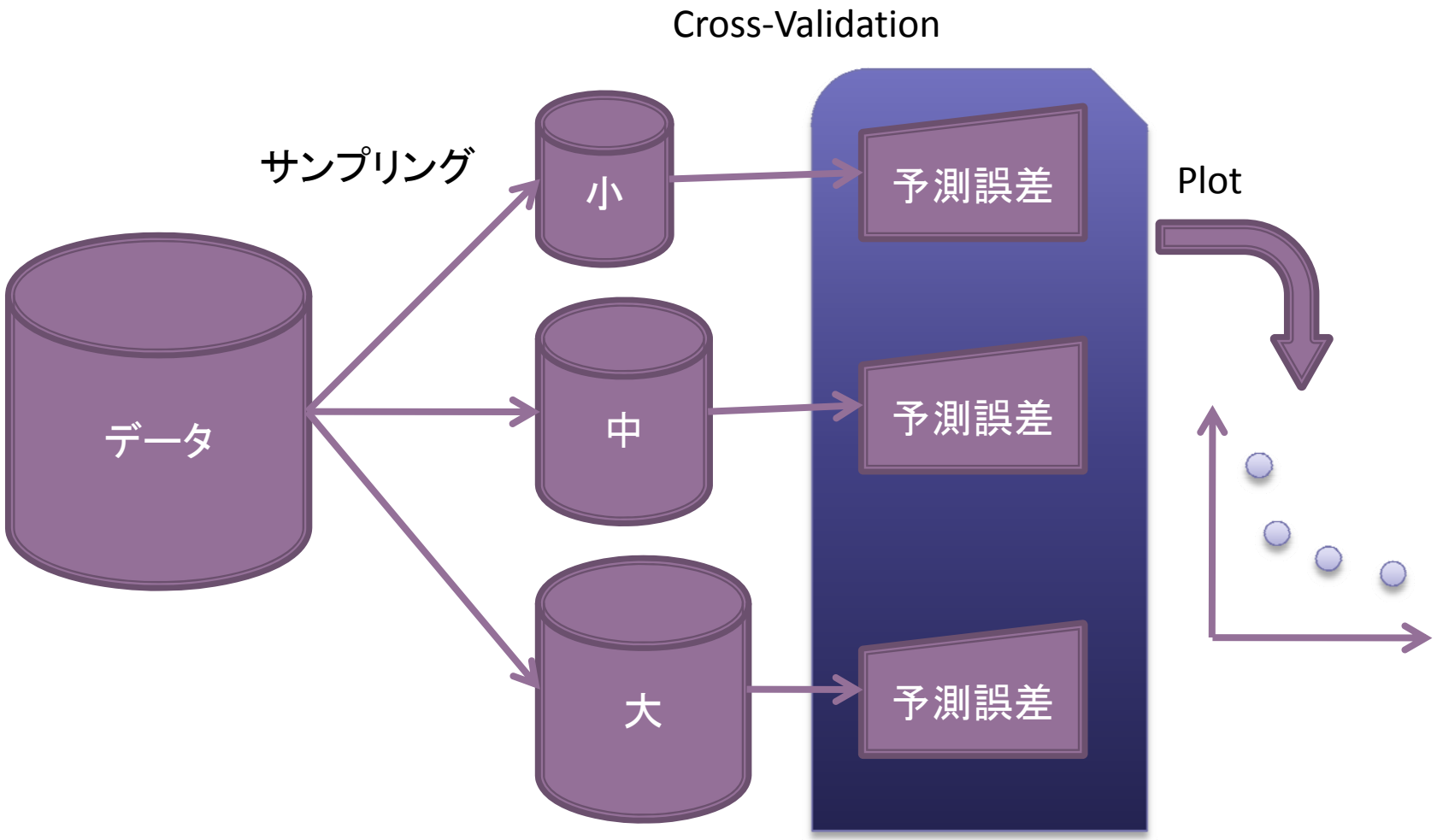
データマイニングはデータが大量になることが多い

- データが大きいと計算に時間がかかる
  - Hadoopなどの分散処理を活用
  - それでも機械学習を適用することは非現実的
    - 計算に数時間～数十日かかることもある
  - サンプリングデータを使いたい
- 全データを使って計算すべき指標とサンプルで十分なもの
  - 全データを使うべき計算
    - “まれ”な値が大きな意味を持つ
    - 例：高額の課金者、長期生存者
      - 集計
      - グラフ、可視化
  - サンプルで十分な計算
    - “モデル”は全体的な傾向を掴むことが目的なのでランダムサンプルで良い
      - 統計的なモデル
- K-sample Plotの紹介（私の博士論文）

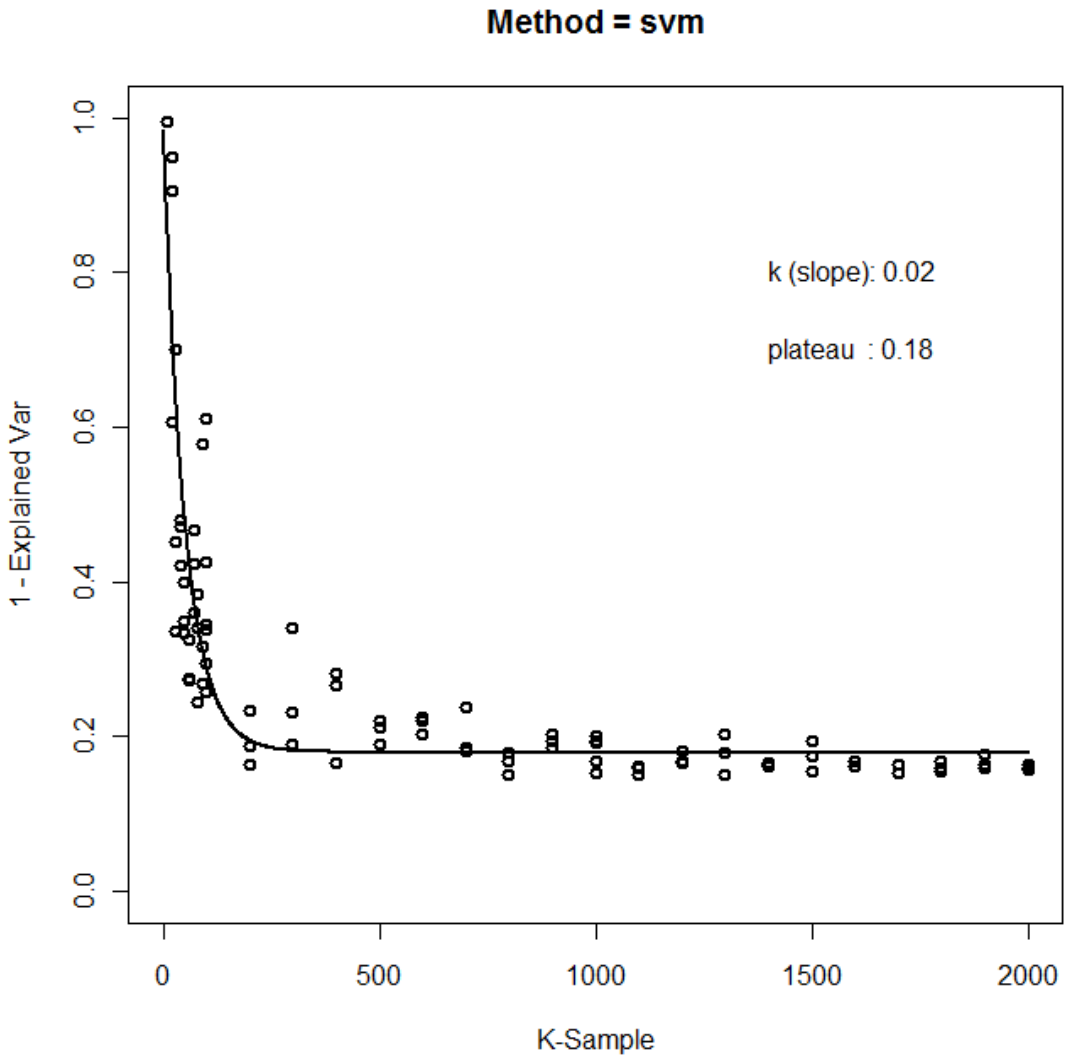
大量データで機械学習を当てはめると時間がかかる

- 1,000,000レコード（100万）のデータでSVMを行う
  - 24時間経っても終わらなかった
- サンプリングを行って予測性能を評価する
- サンプル数を増やしながらエラーをプロットする
- K-sample Plot（K's Plot; Kurahashi Plot）
  - Rのパッケージとして公開しています
  - 「KsPlot」Package
    - （いつかSAS社様にも入れてもらえればいいなぁ）

K's Plotの概念図



K's Plotの例（100万レコードのデータにSVMを適用）



user	system	elapsed
7.35	0.17	7.54

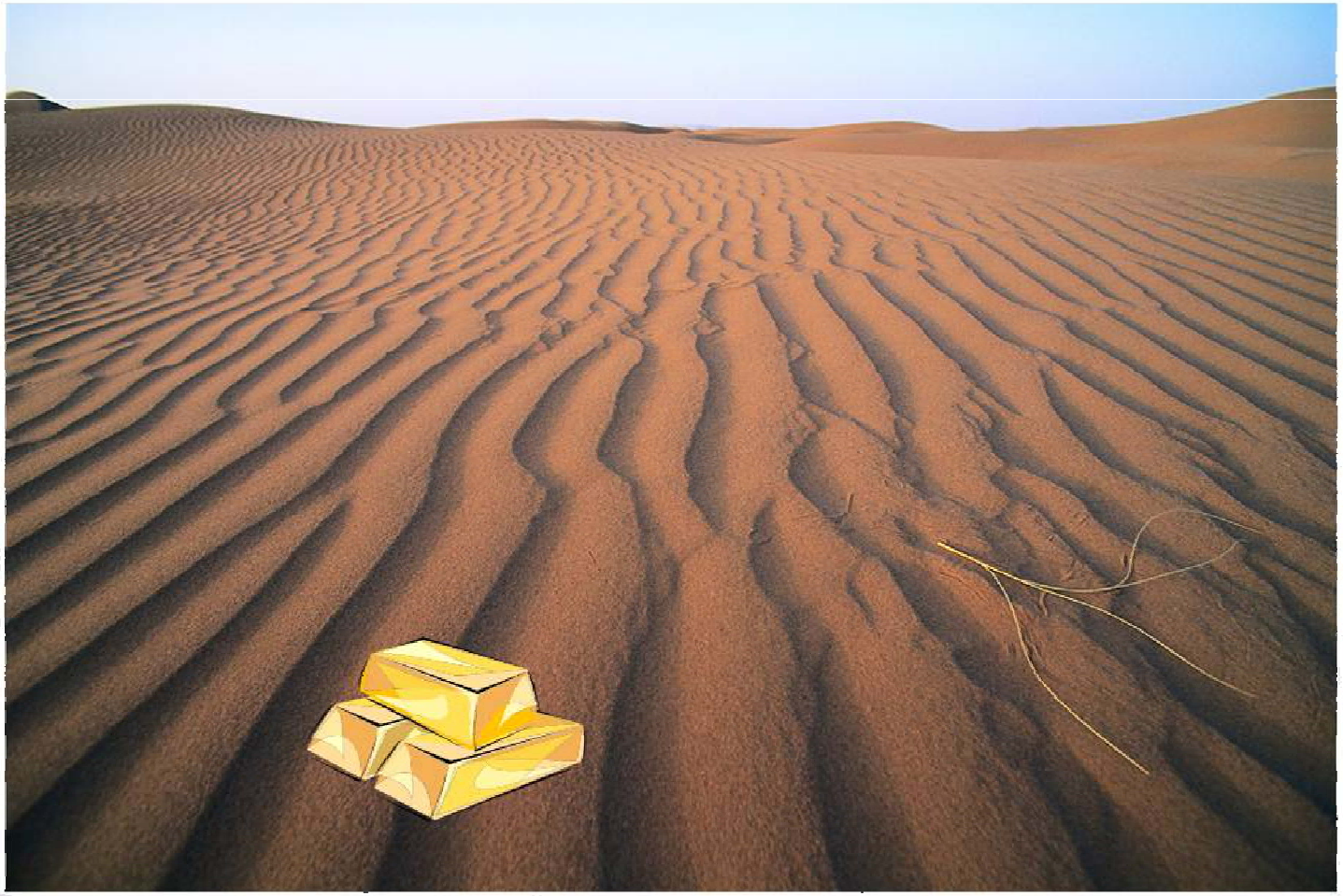
24時間以上
↓
7.5秒

※これはサンプルデータで予測性能を見積もっているので分散処理をしているのではありません

iAnalysis

どうやって大量のデータから金を見つけますか？

How do you find the gold?



http://www.ianalysisllc.com/

iAnalysis