データマイニング
(Data Mining)

神嶌 敏弘

データマイニング

社会の高度情報化 & 情報発信の低コスト化

大量のデータが常に生成されている

記憶媒体の大容量化 & 通信の高速化

膨大なデータの蓄積や流通が可能になった

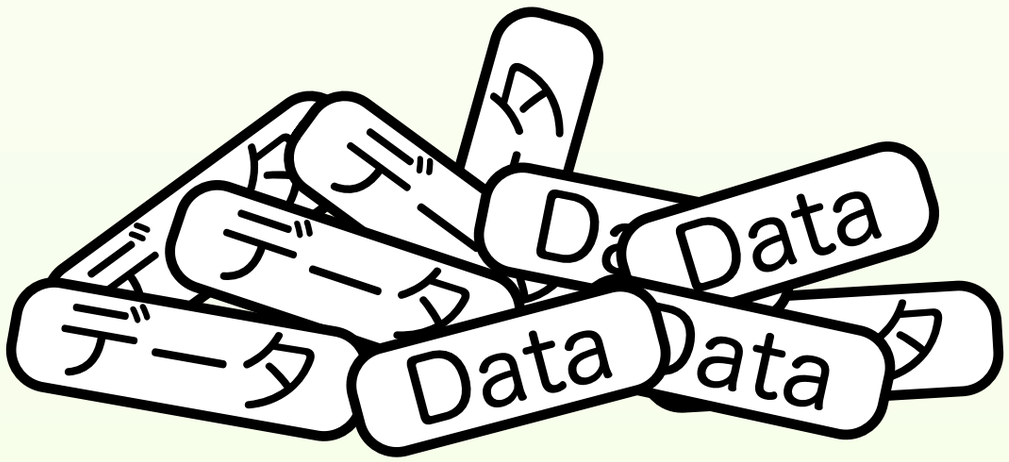
↓

整理されていない膨大なデータの蓄積

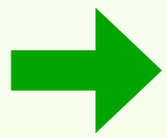
データマイニング

データマイニング

整理されていないデータから，予期されていないが，
再利用可能な知識を掘り起こす(マイニングする)



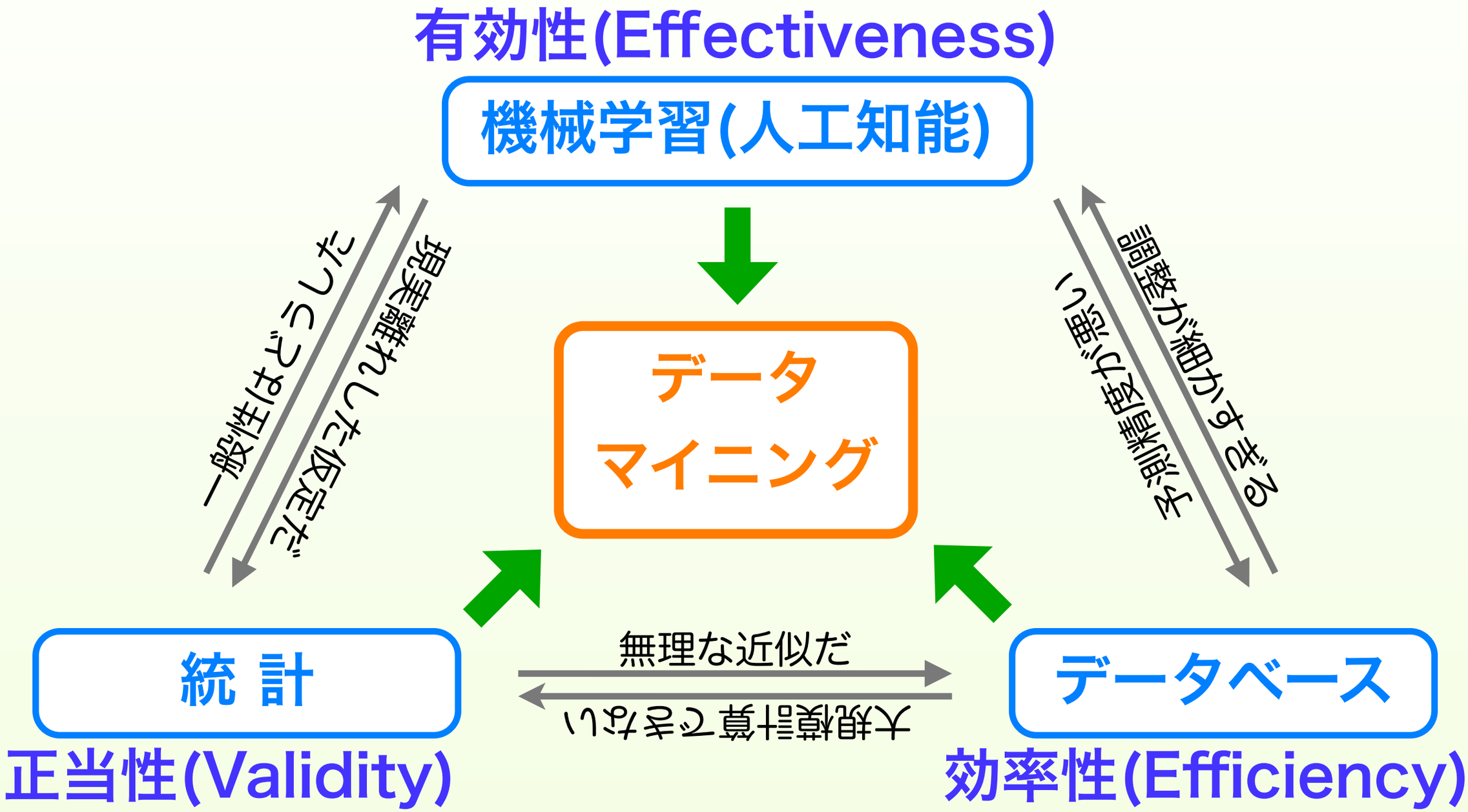
整理されていないデータ



再利用可能な知識

関連技術： 機械学習(人工知能)，統計，データベース
アルゴリズム，分散システム，並列計算機

データマイニング



みんな大切！ バランスよく！

データマイニング

実用性を重視し，データ解析技術の探索的な側面を強調

1. **大規模データ**
   DBアクセス頻度，メモリ効率，計算量は線形
2. **データ収集が計画的や静的ではない**
   はずれ値，欠損値，ストリームデータの処理
3. **新しい種類のデータやパターン**
   時系列，木，グラフ，順序データ，相関ルール
4. **人間による知識の新規性や有用性の判断**
   データの可視化，補助情報や制約の利用

データマイニング

相関ルール (Association Rule)

(頻出パターンマイニングの代表的問題)

Apriori, PrefixSpan, AprioriAll

回帰分析 (Regression Analysis)

線形回帰, ロジスティック回帰, 回帰木

クラス分類 (Classification)

決定木, 判別分析, サポート・ベクトル・マシン

クラスタリング (Clustering)

k-means法, 群平均法, BIRCH

相関ルール

相関ルール Association Rule

$X \Rightarrow Y$

X：前提部 (antecedent)
Y：結論部 (consequent)

X と Y は互いに同じものを含まないアイテムの集合
(アイテム：商品などの「もの」)

X という条件が満たされる場合には，同時に Y という条件が満たされる場合も頻繁に生じる

例：{ 牛乳, パン } ⇒ { 卵 }

牛乳とパン（Xに相当）を同時に買う人は，高い頻度で卵 (Yに相当) を買う

相関ルール

バスケット データ Basket Data

一度のトランザクション(ひとまとめの取引)ごとに，同時に購入された商品の一覧をまとめたデータ

例： T_1 = {牛乳, サンドウィッチ, ハンバーガー}
T_2 = {焼肉弁当, ポテトサラダ}
⋮
T_N = {パスタセット, 牛乳}

トランザクションT_1は，今日の最初の客との取引.
その客の買い物かご(バスケット)には牛乳，サンドウィッチ，ハンバーガーが入っていた

相関ルール

条件Xがトランザクションiを満たす

⇕

Xがトランザクションiの部分集合

例： 条件 X = {牛乳, ハンバーガー}

T_1 = {牛乳, サンドウィッチ, ハンバーガー}

T_2 = {パスタセット, 牛乳}

条件 X ⊆ 取引1 → トランザクション1を満たす

条件 X ⊈ 取引2 → トランザクション2を満たさない

バスケットデータ分析

XとYが共に頻繁に満たされる相関ルールを抽出

相関ルール

利用例 X と Y を組み合わせてセット商品作る

{ 緑茶, ツナおにぎり } ⇒ { タラコおにぎり }
{ 緑茶, ツナおにぎり } ⇒ { コンブおにぎり }

このような相関ルールが共に見つかったとする

緑茶, ツナ, タラコ, コンブおにぎり

このようなセットメニューを作り，単体で買うより少し価格を下げておくと，顧客あたりの購入単価の向上に役立つだろう

相関ルール

利用例

- X と Y を近くに並べて同時購入を促す

  {インスタントラーメン} ⇒ {チャーシュー}

インスタントラーメン売り場にチャーシューを並べると同時購入が増えるだろう

- Xの販売数が増加する場合にYの在庫を増やしておく

  {牛乳} ⇒ {パン}

牛乳の特売をすると，牛乳の販売数が増えると，パンの販売数も増えると予測される
よって，パンの仕入れ数を増やしておく

相関ルール

Apriori

IBMのRakesh Agrawalが中心になり開発
大規模なバスケットデータから次の条件を満たす相関ルールをすべて発見する

- XuY を満たすトランザクション数は多い
- Xが満たされるときYも高頻度で満たされる

PrefixSpan，AprioriAll

時間的な変化も考慮する Apriorioriアルゴリズムの拡張

例：「以前デジタルカメラを購入」し，その後「ケースを購入」した顧客は，さらにその後に高い頻度で「メモリカード」購入する

回帰分析

回帰分析 Regression Analysis

入力： 事例 (y_i, x_i) とは対象を表す変数 X の値が x_i の場合に，結果を表す変数 Y の値が y_i であったという具体例

事例集合 $(y_1, x_1), (y_2, x_2), \cdots, (y_N, x_N)$

様々な対象 X に対応する Y の値との組である事例をN個集めたもの

出力： 変数 X と Y との関数関係 f

$$Y = f(X)$$

X：説明変数，独立変数，属性，特徴

Y：被説明変数，従属変数

f：回帰直線(曲線)

回帰分析

利用例

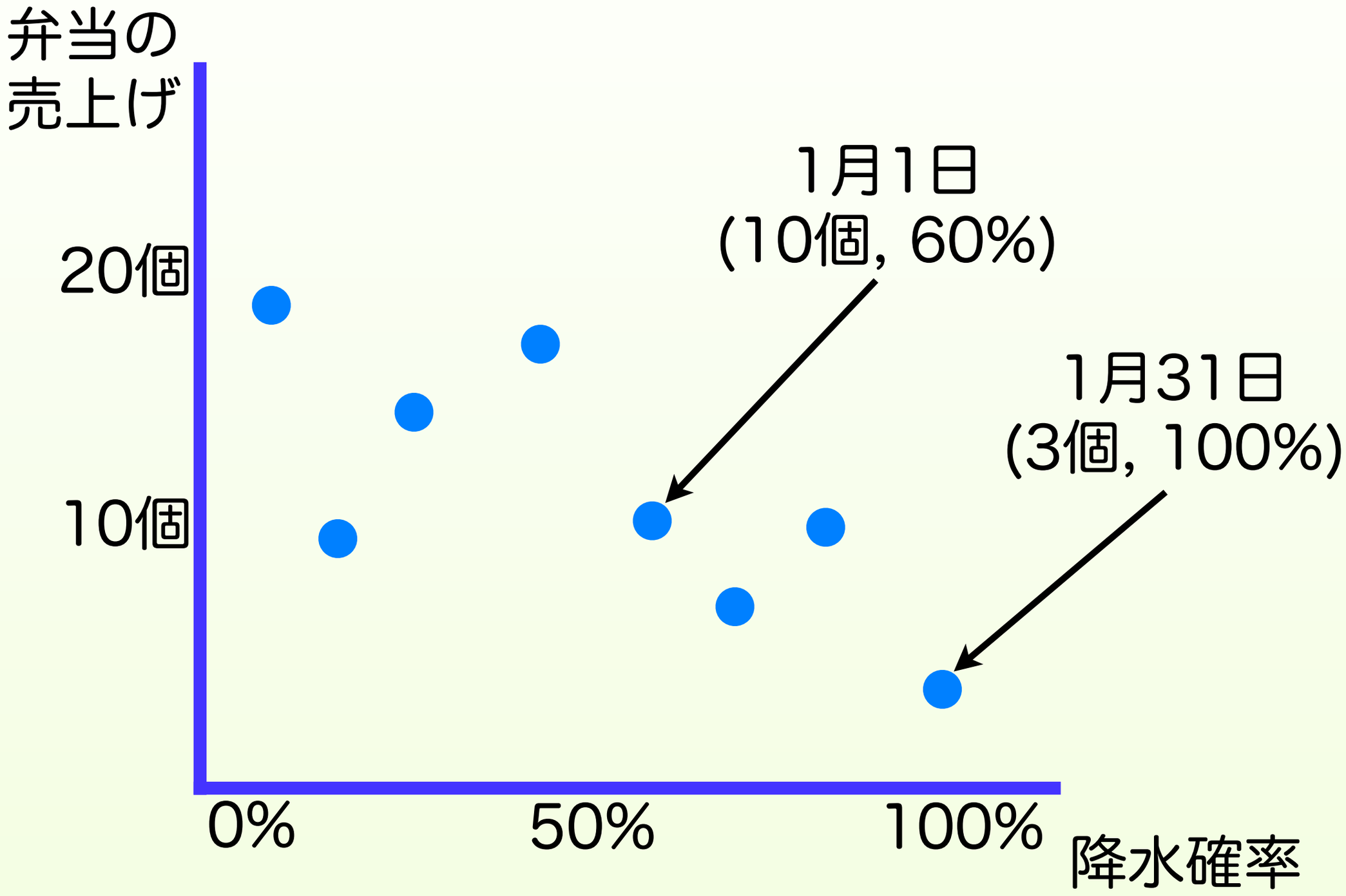
天候
- 晴 → 弁当が売れる → 在庫不足
- 雨 → 弁当が売れない → 売れ残り

明日の降水確率から弁当が売れる数を予測する

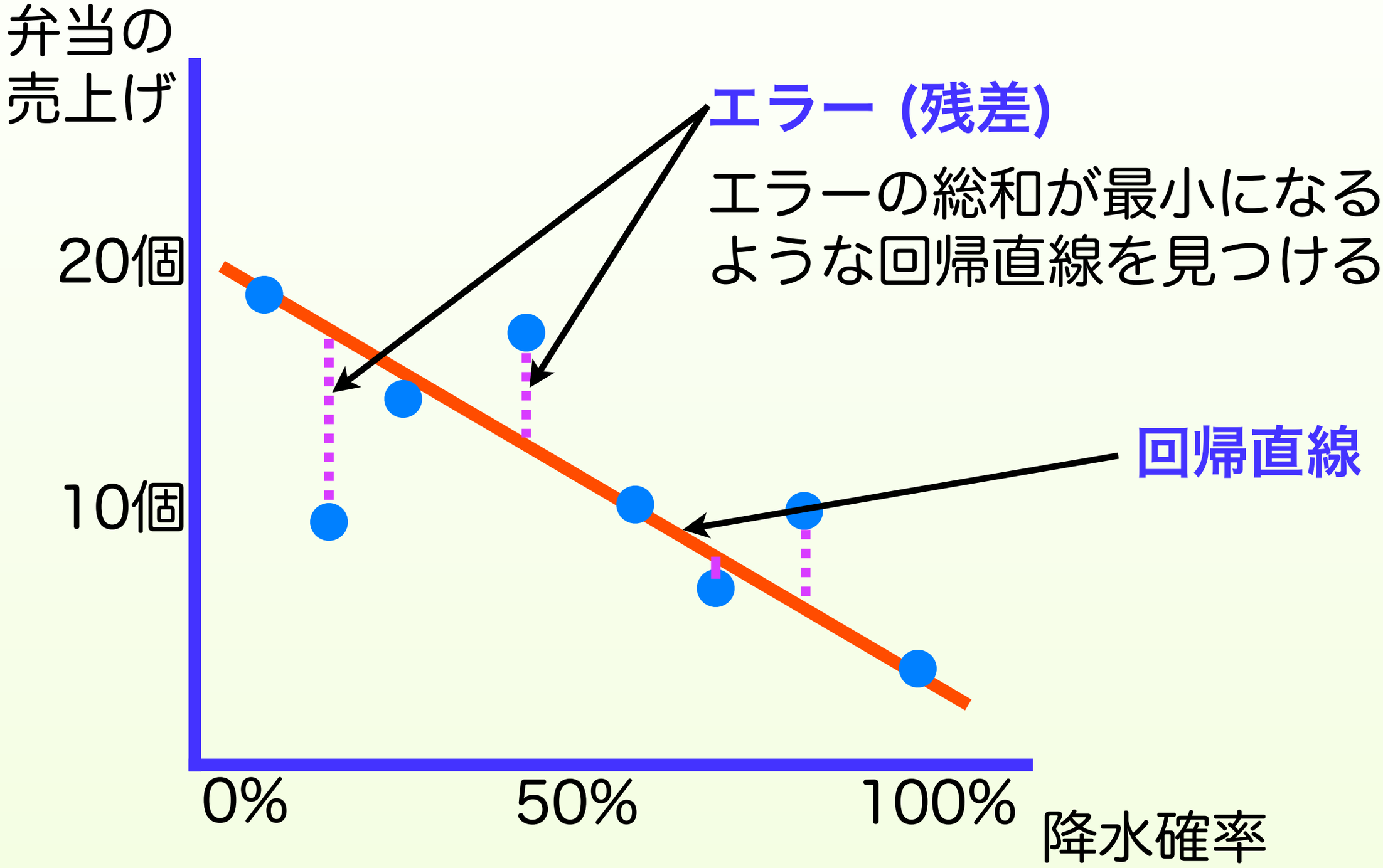
説明変数　X：明日の降水確率
被説明変数 Y：弁当の売れる数

事例集合：過去の降水確率と売り上げのデータ
1月1日 (10個, 60%)　1月2日 (20個, 0%)
・・・・ 1月31日 (3個,100%)

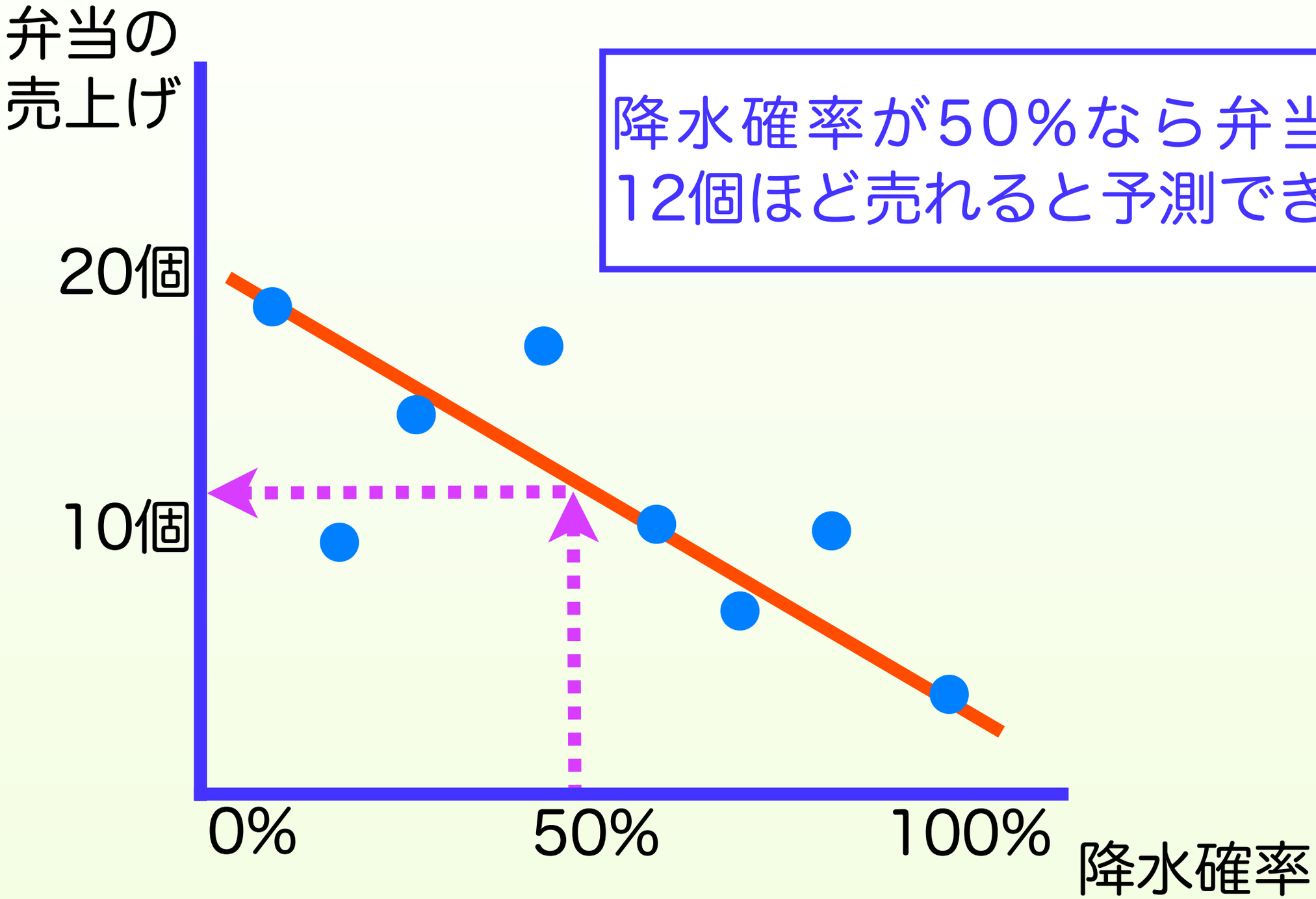
回帰分析



回帰分析


弁当の
売上げ
20個
10個
0%
50%
100%
降水確率
エラー (残差)
エラーの総和が最小になる
ような回帰直線を見つける
回帰直線

回帰分析



降水確率が50%なら弁当は
12個ほど売れると予測できる

回帰分析

回帰直線

$$Y = -0.2X + 3$$

説明変数　X：明日の降水確率
被説明変数 Y：弁当の売れる数

弁当の売上げは降水確率以外のの要因にも影響される

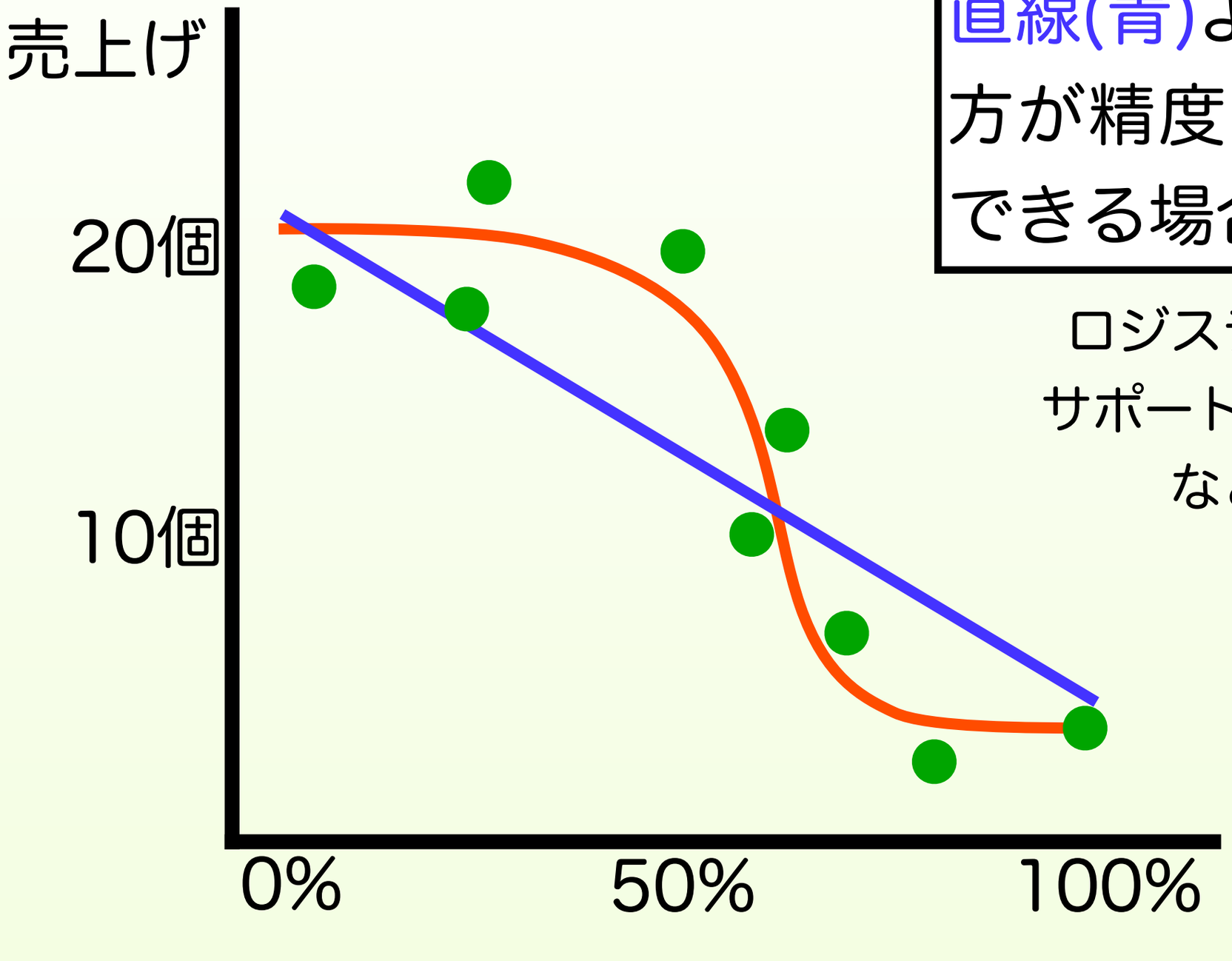
説明変数　X_1：明日の降水確率
　　X_2：平日は1，土日祝は0
被説明変数 Y：弁当の売れる数

$$Y = -0.2X_1 + 2.2X_2 + 3$$

統計では説明変数が2個以上なら**重回帰分析**と呼ぶ

回帰分析

非線形回帰



直線(青)より曲線(赤)の方が精度の高い予測ができる場合もある

ロジスティック回帰
サポートベクトル回帰
などの手法

クラス分類

クラス分類 Classification

入力： 事例 (c_i, x_i) とは対象を表す変数 X の値が x_i の場合に，変数 C の表すクラスが c_i であったという具体例

事例集合 $(c_1, x_1), (c_2, x_2), \cdots, (c_N, x_N)$

様々な対象 X に対応するクラス C の状態の事例をN個集めたものが事例集合

出力：分類規則(識別規則)

対象 X に対応するクラス C を予測するのが分類規則

クラス分類

クラス

事前に定めた有限個の集合中のいずれかの値をとる

例： おすすめ商品に {関心あり, 関心なし}

天気のクラス {晴, 曇, 雨, 雪}

競馬で1位の馬のクラス {1番～15番}

利用例

スーパーのポイント会員にバーゲンの案内状を送る

バーゲンに来ない客に送ると余分な郵送費がかかる

できるだけ来てくれそうな客を選びたい

例：

対象　　X：会員
クラス C：バーゲンに {来る, 来ない}

クラス分類

対象 X は幾つかの **属性 (特徴)** で表現する

これらの属性を並べたものを **属性ベクトル** という

$$i\text{ 番目の対象 } X_i = \langle X_{i1}, X_{i2}, \cdots, X_{im} \rangle$$

例： ポイント会員の属性
バーゲンに来る可能性に影響しそうなものを選ぶ

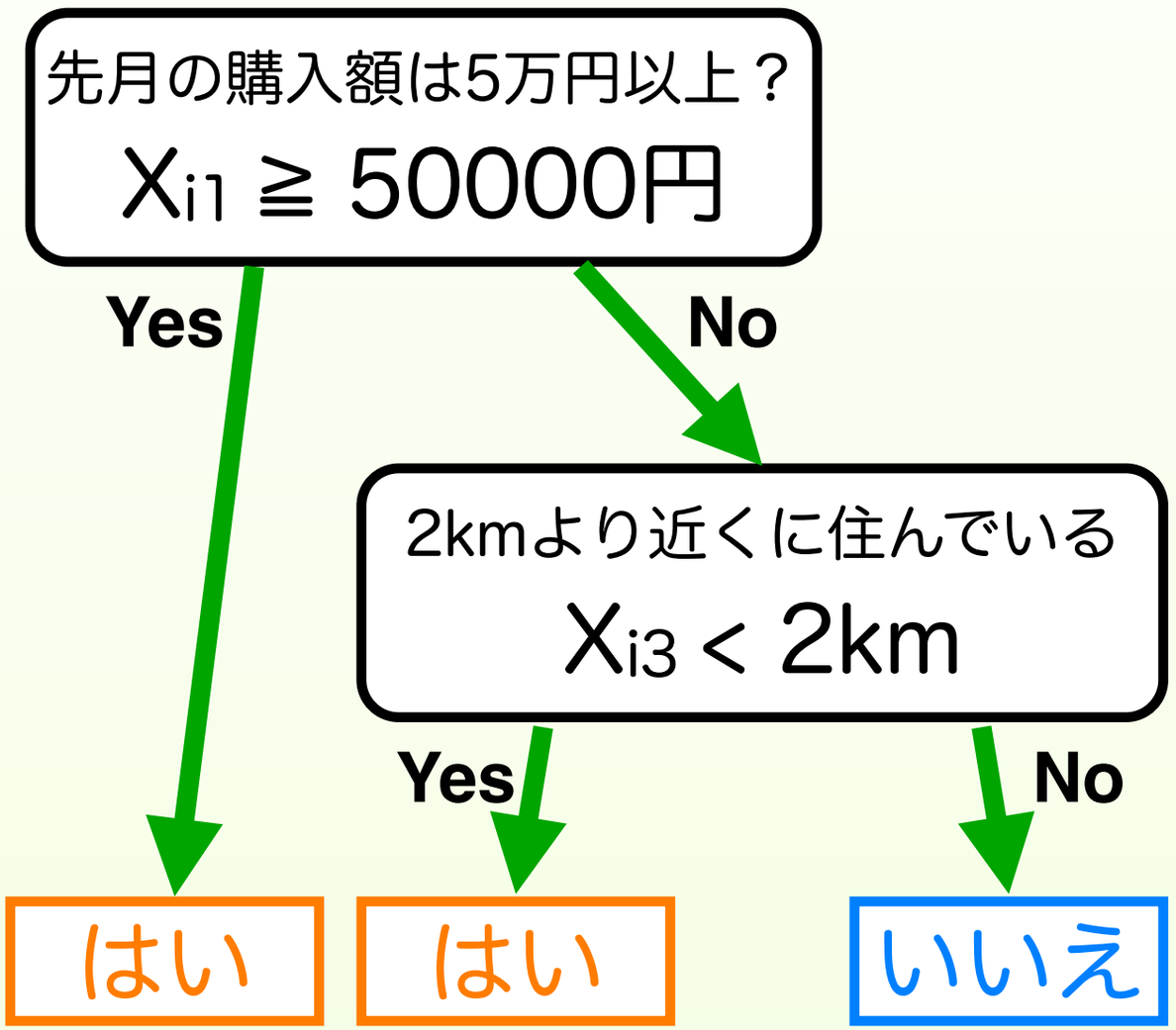
第1属性 X_{i1}：先月の購入額
第2属性 X_{i2}：先月の来店回数
第3属性 X_{i3}：店舗から住所までの距離

例： 会員3番 $x_3 = \langle$4万円, 25回, 1.5km$\rangle$
前回のバーゲンでは来店 クラス c_3 = はい
事例3 (c_3, x_3)=(はい, $\langle$4万円, 25回, 1.5km$\rangle$)

クラス分類

分類規則の例

決定木

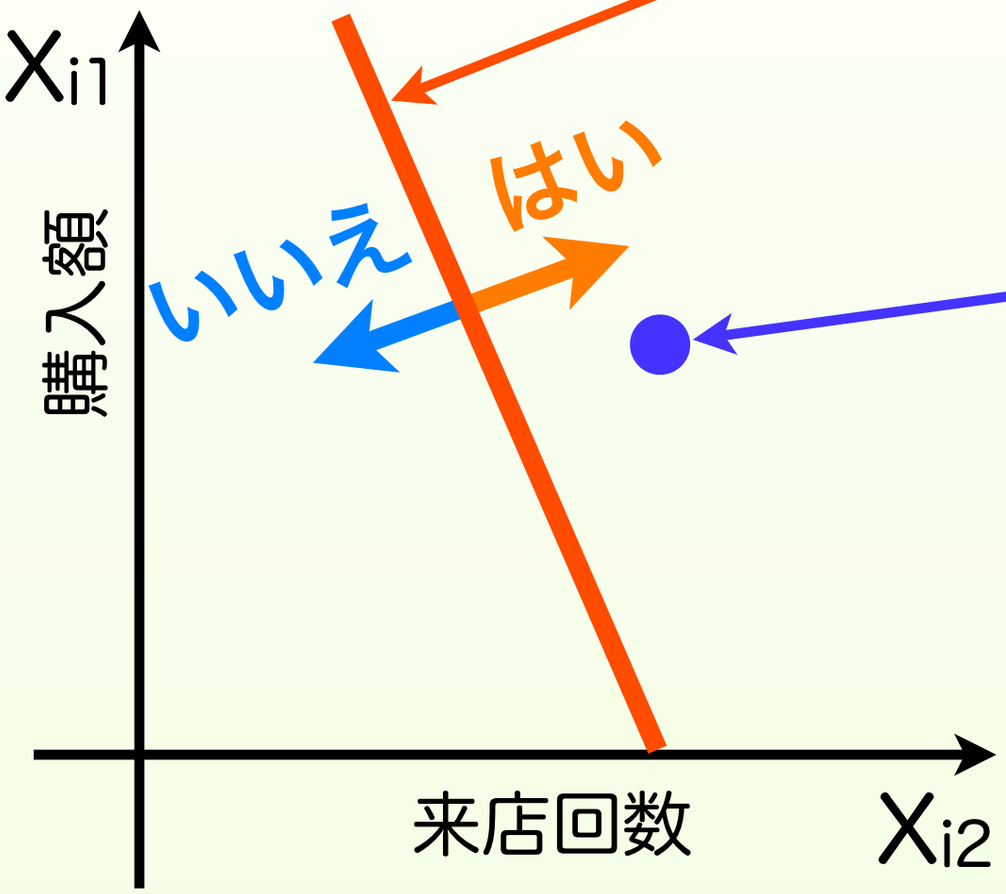


例：購入額が5万円未満で2kmより近くに住んでいる会員はバーゲンに来店する

クラス分類

分類規則の例

識別境界面



会員5番は来店する
(X_{i1}=6万, X_{i2}=20回)

識別境界面のどちらに対象があるかでクラスを決定

クラス分類

分類規則の主な学習手法

- **Fisher判別分析**

  最も古くに考案された基本的な方法

- **決定木学習 (ID3, CARTなど)**

  結果の解釈が容易, 比較的高速

- **サポートベクトルマシン**

  高次元で有利, 複雑な分類規則を学習可能

- ***k*-近隣法**

  分類規則を作らず, 事例そのものを使って分類

- **ベイズ推定**

  理論的な背景が堅固, 計算は大変, ノイズに強い

回帰分析とクラス分類の相違点

回帰分析		クラス分類
事例：(y_i, x_i)	⇔	事例：(c_i, x_i)
説明変数 x_i の値に対応する被説明変数 y_i の値		属性 x_i の値に対応するクラス c_i の値
↓		↓
被説明変数 y_i の値を予測する回帰曲線	⇔	クラス c_i の値を予測する分類規則

被説明変数は数値変数だがクラスはカテゴリ変数

数値変数：実数や整数などの数値をとる

カテゴリ(名義)変数：有限集合中のいずれかの値をとる

クラスタリング

クラスタリング Clustering

クラスター分析 (Cluster Analysis) とも呼ぶ

入力： 対象集合 $x_1, x_2, \cdots, x_N$

対象 X は属性ベクトルで記述
任意の対象X_1とX_2の似ている度合いを数値化した類似度

出力： クラスタ と呼ぶ次のような部分集合に分割

内的結合(internal cohesion)
同じクラスタ内の対象は互いに似ている

外的分離(external isolation)
違うクラスタにある対象は似ていない

クラスタリング

利用例

スーパーのポイント会員を，その購買の傾向が似ているクラスタに分割する

i 番目の会員 X_i を次のような属性ベクトルで表現

第1属性 X_{i1} ：先月の弁当・総菜の購入額
第2属性 X_{i2} ：先月の清涼飲料の購入額
⋮
第m属性 X_{im}：先月の野菜類の購入額

例： 会員3番 x_3 ＝〈3000円, 1200円, …, 0円〉

クラスタリング

見つかるクラスタの例

- **クラスタ1**：主に生鮮食料品を購入する顧客
- **クラスタ2**：主に中食製品を購入する顧客

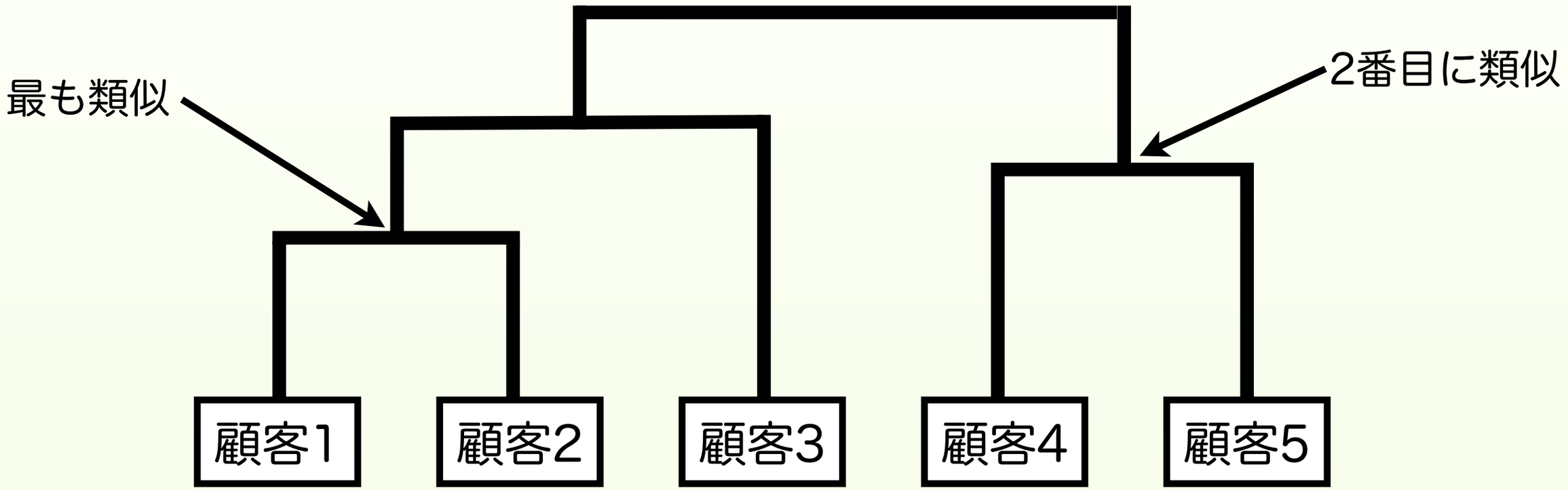
見つかったクラスタの利用例

- バーゲンなどでまとめ買いをするので，ダイレクトメールなどの販促策はクラスタ1の顧客を中心に
- その場で購入するものを決めるクラスタ2の顧客にはセットメニューによる客単価の向上などを中心に

クラスタリングは，データを大まかに要約し，その全体像を把握するための探索的手法

クラスタリング

階層的クラスタリング



最も似ている顧客から順にクラスタにまとめる

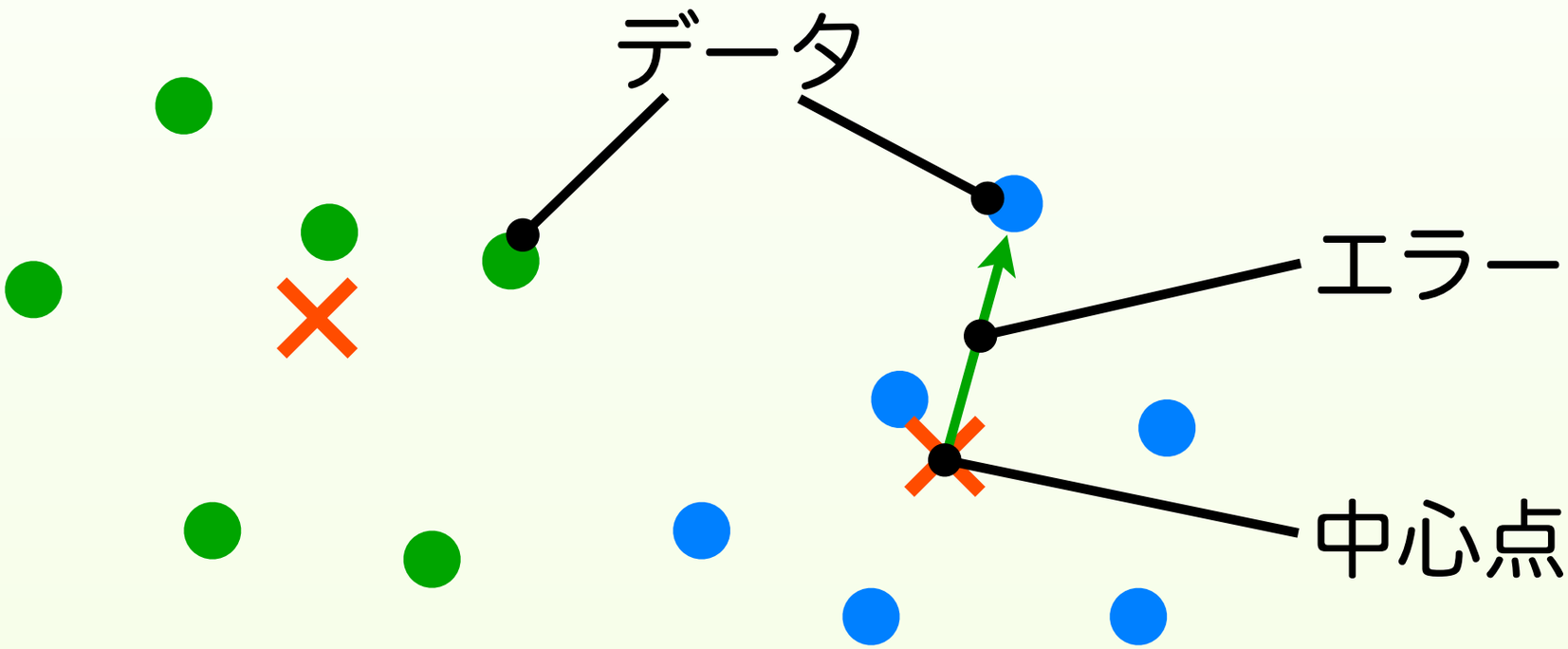
このような図を デンドログラム

代表的手法： 最短距離法，最長距離法，群平均法，Ward法

クラスタリング

分割最適化クラスタリング (非階層的クラスタリング)

クラスタの良さの基準を最適化するように分割



K-means法 代表的な分割最適化手法

中心点とデータの間のエラーが最小になるように中心点とデータの分割を決める

クラスタリング

最新の研究課題

- **大規模データへの対応**

  データの管理を工夫 (BIRCHアルゴリズム等)

- **時系列やグラフなどのクラスタリング**

  属性ベクトル以外で記述されたデータを扱う

- **スペクトラルクラスタリング**

  類似度に基づく分割の新手法

- **カーネルの利用**

  複雑な形状のクラスタの抽出など

- **半教師ありクラスタリング**

  分割を決めるときにヒントを利用

教師あり学習と教師なし学習

教師あり学習 回帰分析，クラス分類

データに加えて予測すべきものの正解も与える

対象 x_i に対応する，クラスの値 c_i や被説明変数の値 y_i が必要

教師信号 データと共に与えた，クラスや被説明変数の具体例

教師なし学習 クラスタリング，相関ルール

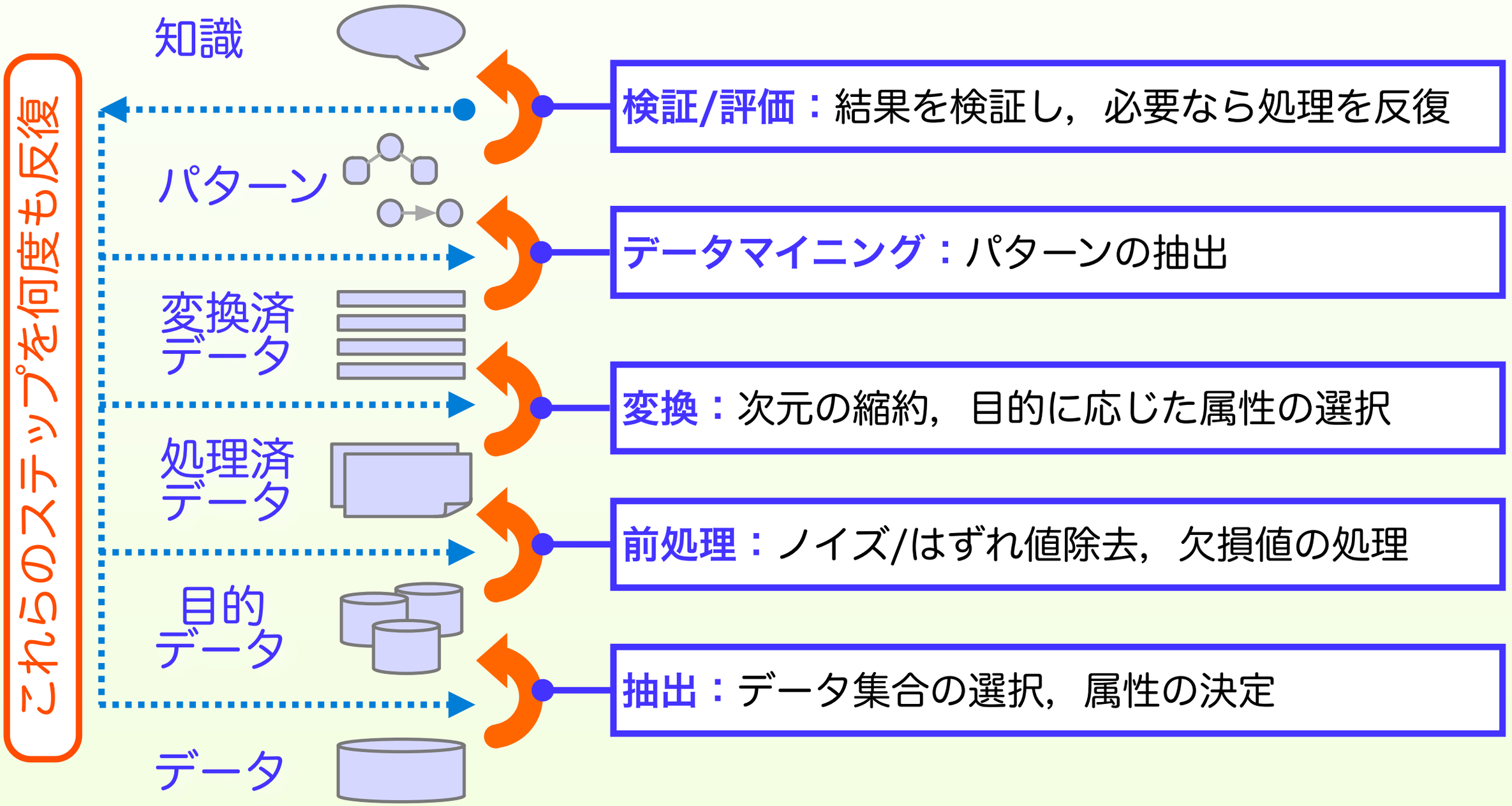
データだけを与える

分類する対象 x_i やトランザクション T_i のデータだけ

知識発見の過程

知識発見(＝広義のデータマイニング)

KDD Process: Knowledge Discovery and Data Mining



知識発見の過程

抽出

例：スーパーマーケットの場合

データ集合の選択 どのデータベースを使うのか

顧客との取引(POS)データ，ポイント会員登録データ，
天候のデータ，周辺の地図データ ……

属性の決定 データベースのどの項目を使うのか

クラス分類の例題で，自宅までの距離は関係ありそうだが，電話番号はあまり関係ないだろう

あとでデータの削減をするので，この段階では多めにデータを保持しておく方がよい

知識発見の過程

前処理

ノイズ/はずれ値の除去 非現実的な値の除去

計測の誤り：計測機器の誤差
誤入力：位の誤り，データの変換ミス
意図的な誤り：年齢のサバよみ，プライバシー問題

欠損値の処理 不足したデータの補完

欠損値：入力されていないデータ

- 同じ項目の，入力済みの値の平均値等を使う
- データ解析に影響を及ぼさない値にする

知識発見の過程

変換

次元削減 主成分分析

データを関数で変換して少ない属性数で表現

目的に応じた属性の選択 ラッパー法とフィルター法

問題の解決に関連がありそうな属性以外は削除

次元の呪い

属性が多すぎると適切なパターンを発見できなくなる

- データの分布が均一に近づき，データ間の類似の度合いが均一になる
- パターンがあるかどうかを，少ないデータ量では検証できなくなる

知識発見の過程

検証/評価 結果を検証し，必要なら処理を反復

データ，前処理，属性，データマイニング手法の適切さを，解析している問題についての知識を使って検証

例：相関ルール抽出

{レトルトカレー} ⇒ {洗濯用洗剤}

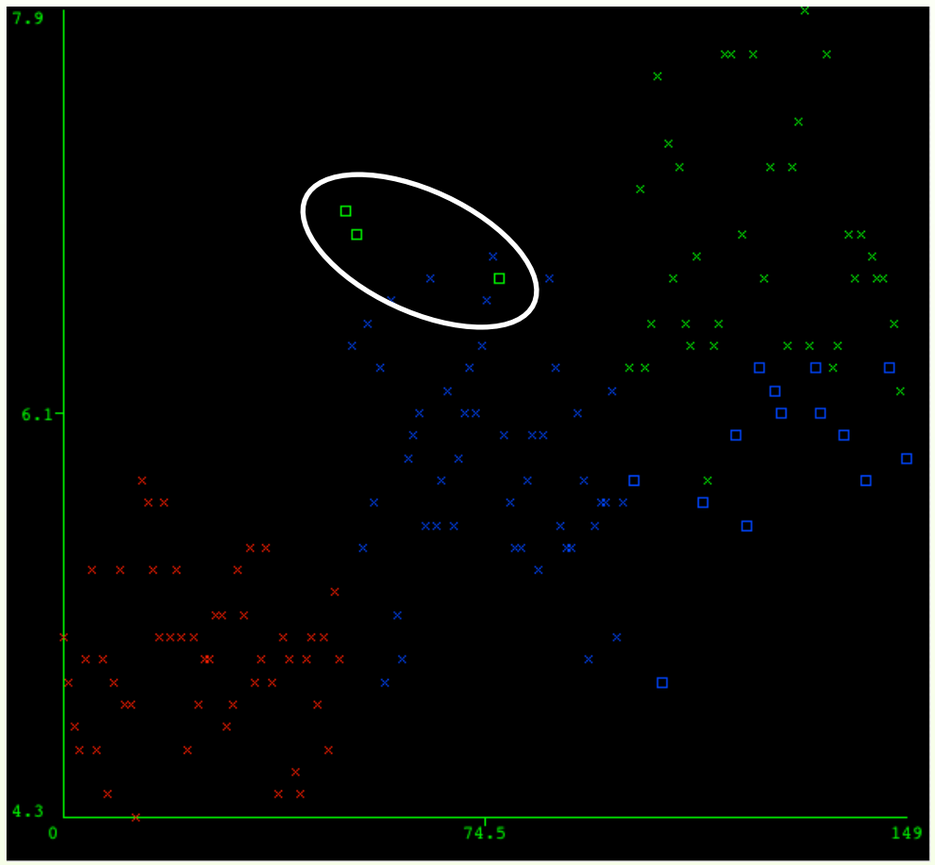
× カレーを食べると服が汚れるから

○ たまたま，カレーと洗剤を同じ日に特売した

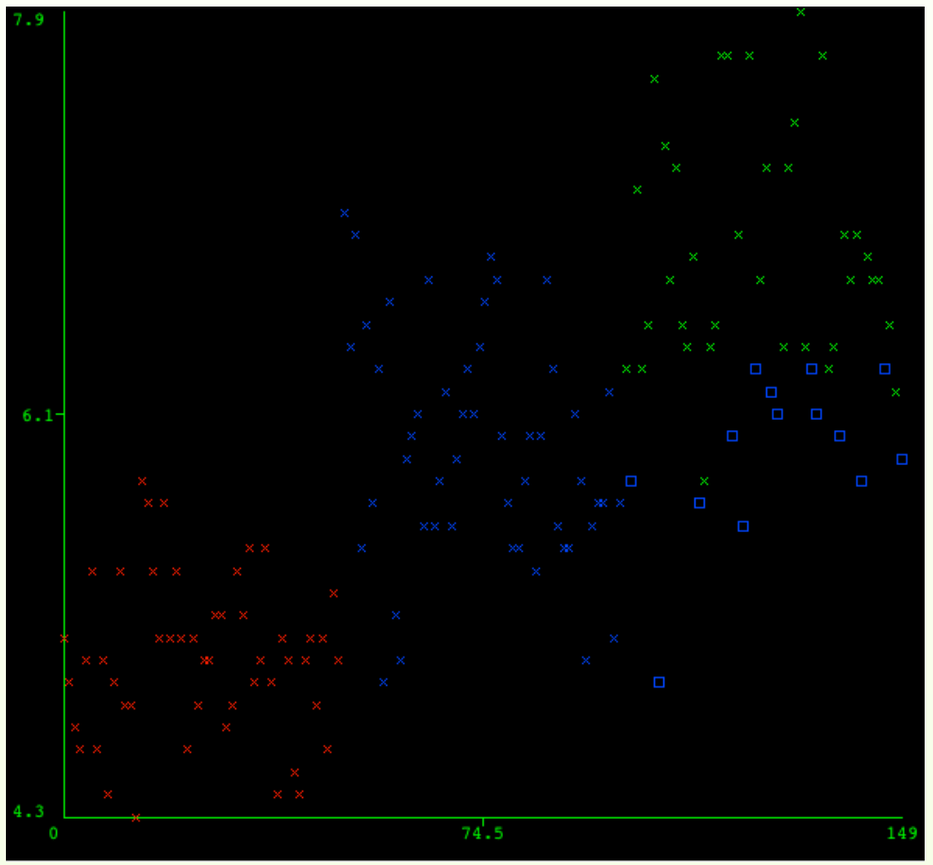
見つけた知識の合理性は必ず検証

知識発見の過程

例：データマイニング手法の「クセ」に由来する場合
同じデータでも手法が違うと……



K-means法



混合ガウスモデル+EMアルゴリズム

「似ている」ことの基準が違う

データマイニング手法の解析の前提を理解すること

まとめ

- **データマイニングとは？**

  整理されていデータから，予期されていないが，再利用可能な知識を見つける

- **代表的なデータマイニング手法**

  相関ルール：X が起きるときには Y も起きやすい

  回帰分析：X の属性から，数値変数 Y を予測

  クラス分類：X の属性から，そのクラス C を予測

  クラスタリング：似ているものどうしをまとめる

- **知識発見の過程**

  見つけた知識は，無条件に受け入れずに必ず検証し，より妥当な知識を見つけるよう試行錯誤

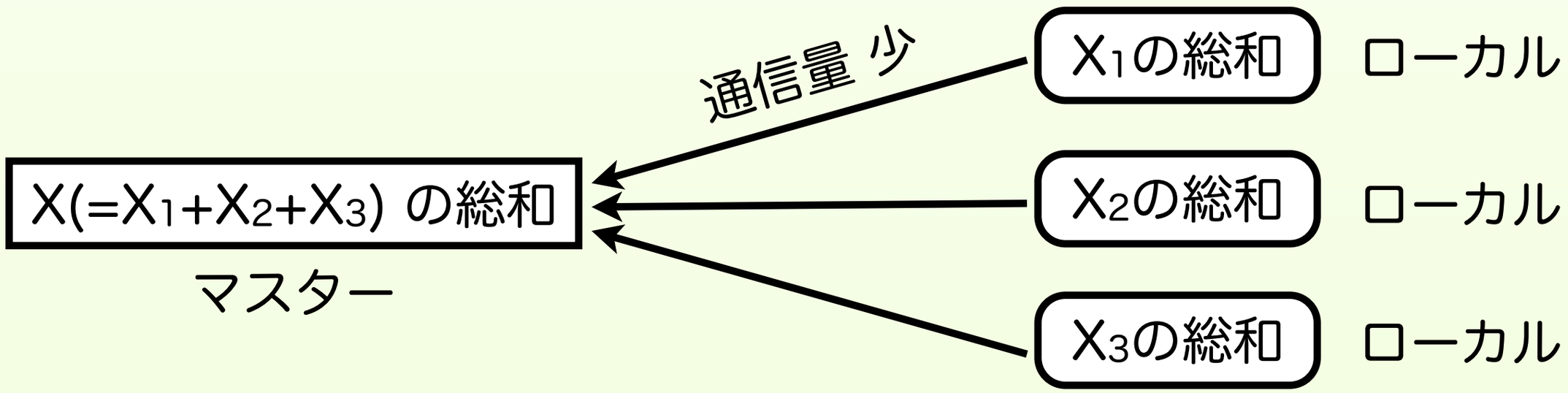
分散データマイニング

- 複数の計算機を利用して計算を高速化
- データが地理的に離れた場所 (支社・部署ごと) で保持されている

↓

分散データマイニング

- データ全体を分割してローカル計算機で保持
- 担当データの主な計算はローカル計算機内で実行
- ローカル計算機間の通信量は少なく保つ



プライバシー保護データマイニング

一見，無関係な情報も……　　プライバシー侵害の問題

性別 + 誕生日 + 5桁郵便番号 ➡ 87%の米国人は一意に特定可能

個々のデータ	➡	全体の傾向
プライバシー ✗		プライバシー ○

個々のデータは秘密のままマイニング
プライバシー保護データマイニング

水平分割型

	特徴1	特徴2	特徴3
Aさん			
Bさん			
Cさん			

個人間でデータは秘密

垂直分割型

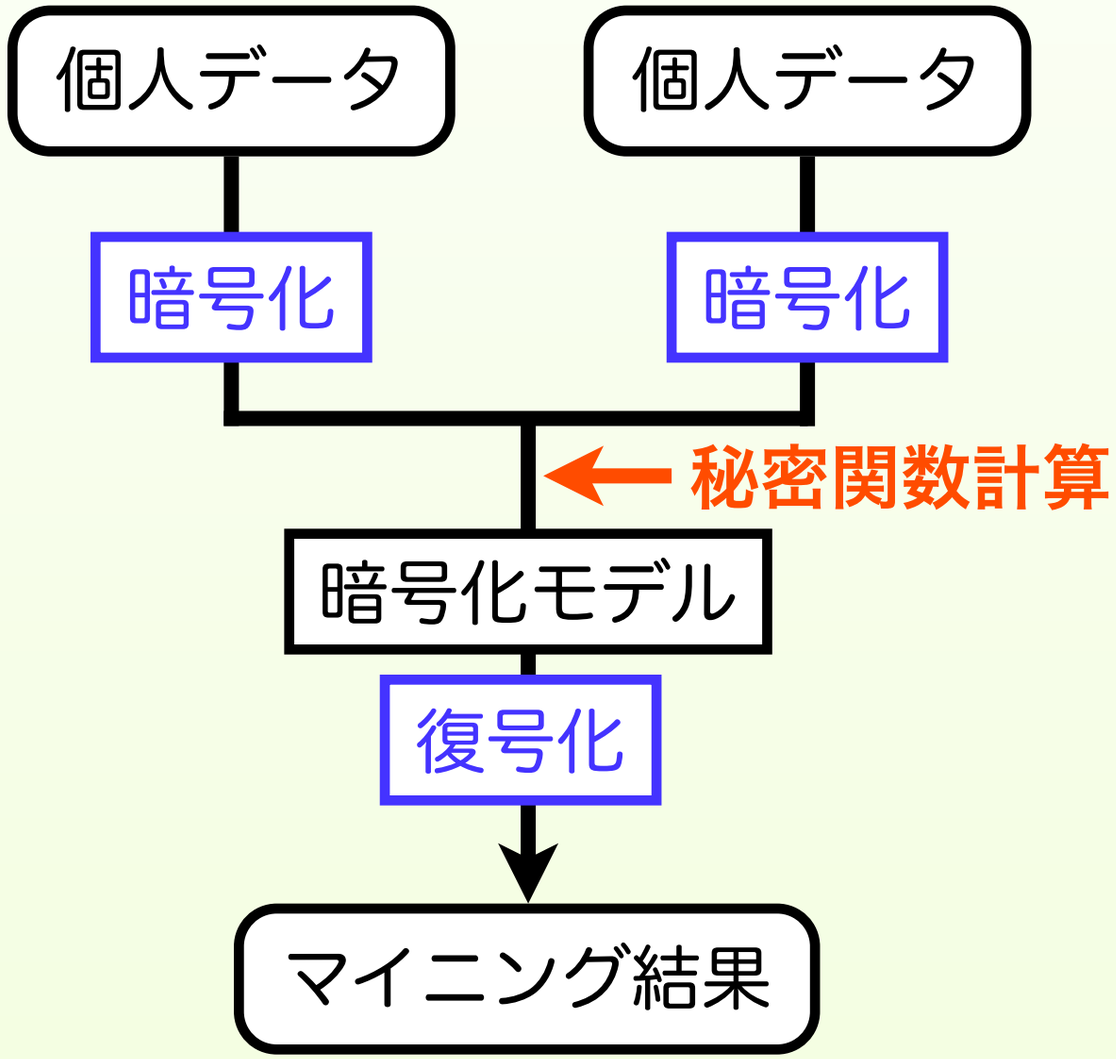
	特徴1	特徴2	特徴3
Aさん			
Bさん			
Cさん			

各人の異なる特徴が秘密

プライバシー保護データマイニング

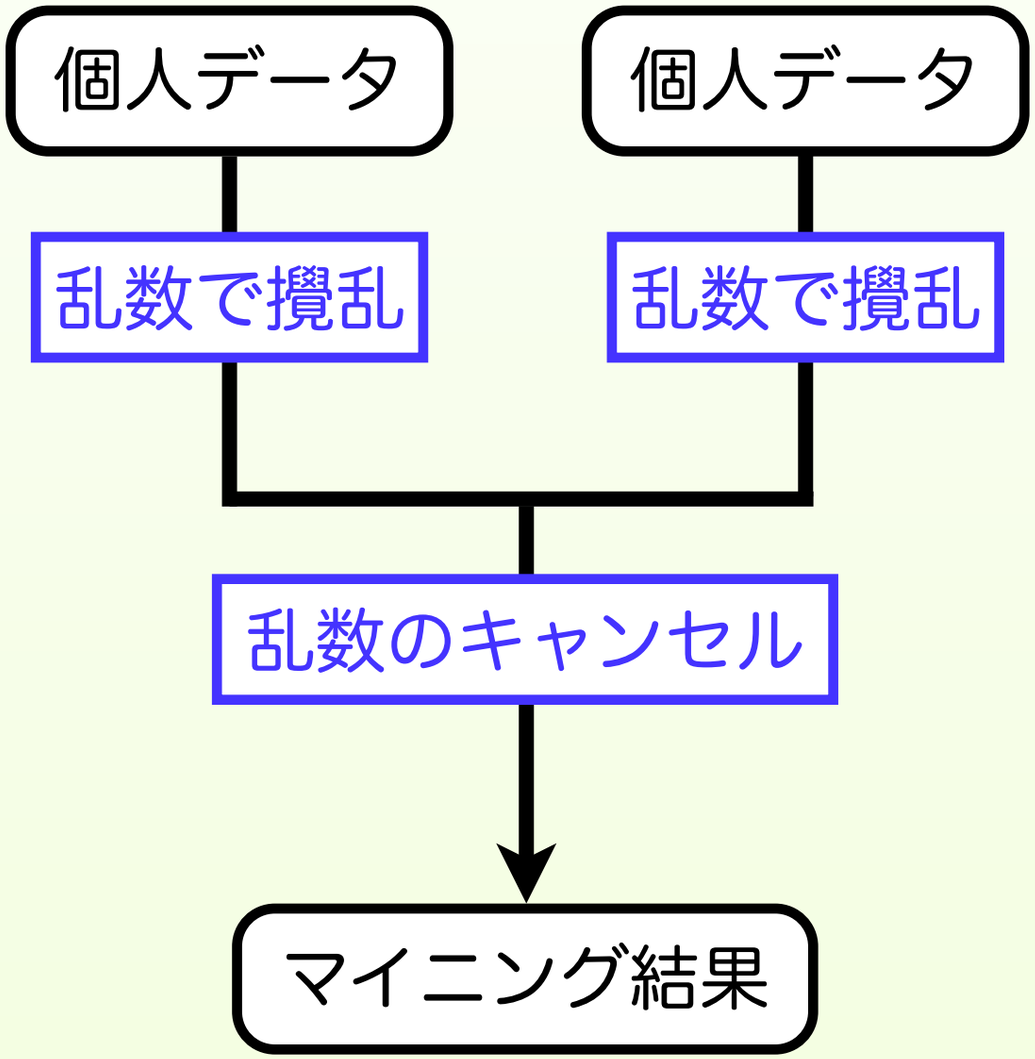
秘密関数計算

データを暗号化し，暗号化したまま秘密に結果を計算



ランダム化

計算過程で，互いに打ち消し合うような乱数で元データを攪乱



時系列データ

時系列データ：時間軸上で反復的に計測された値の系列

例：株式市場分析，経済・販売予測，負荷予測，生産・品質管理，気象・地質データ，科学・技術実験，医療

処理の種類

- **類似度検索，索引付け**
  クエリ時系列 Q と類似度に基づき，時系列DBから類似した系列を見つける
- **クラスタリング**
  類似度を元に，時系列DB中の時系列を類似したグループに分割
- **クラス分類**
  ラベルのない時系列 Q を，事前に定めたクラスに分類
- **セグメンテーション，トレンド分析**
  長さ N の時系列 Q を，K≪N 個の部分に分ける

Webマイニング

Webマイニング：Webの文書やサービスから情報を自動的に発見したり抽出したりするためのデータマイニング技術の利用

- **Web内容マイニング (Web content mining)**

  Webの内容/データ/文書から有用な情報を発見

  自然言語処理や情報検索の技術がほぼそのまま適用される

- **Web構造マイニング (Web structure mining)**

  Webのリンク構造に内在するモデルの発見

  リンクの参照関係から，Webに反映された社会コミュニティの発見

- **Web利用マイニング (Web usage mining)**

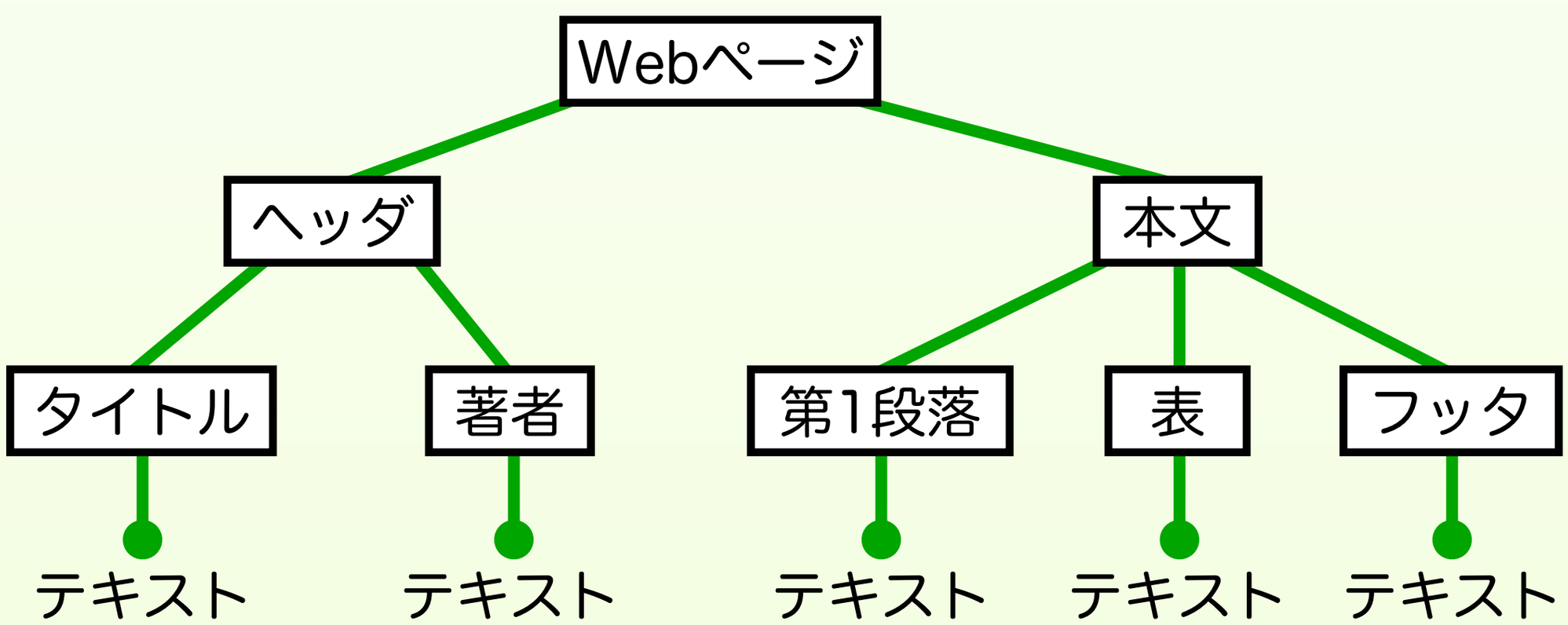
  Web閲覧者のセッションや行動で生成されたデータの理解

  Webの閲覧ログを利用した利用者行動の予測

半構造データ

半構造データ：構造のないテキストを構造化された枠組みに格納したもの

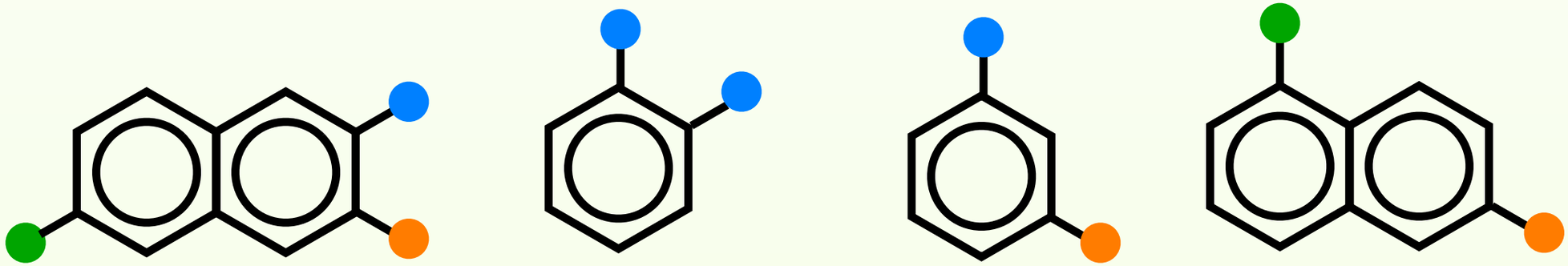
例：Webページ，XMLデータ (タグで表された構造とテキスト)
定型化されたテキスト (タイトルやアブストなどの構造にテキストが格納されている)



グラフマイニング

グラフマイニング：グラフで表現されたデータからのマイニング

例：回路，化合物，タンパク構造，生物学的ネットワーク，社会ネットワーク，Web，ワークフロー，XMLデータ



化合物のグラフ：結合が辺，分子や基がノード
このグラフの中から，頻出する部分グラフを抽出
➡ 特定の性質・薬効をもつパターンに相当

- グラフは同一性の判定が難しいので特殊な索引付け技術が必要
- カーネルを使う方法もある

参考文献

- 朱鷺の杜Wiki：http://ibisforest.org/index.php?FrontPage
  データマイニング・機械学習の情報をまとめています
- C.M.Bishop “Pattern Recognition and Machine Learning” Springer (2006)
  日本語版「パターン認識と機械学習 上下 - ベイズ理論による統計的予測」シュプリンガー・ジャパン (2007-2008)
  http://ibisforest.org/index.php?PRML
  最近の機械学習手法を俯瞰できる本
- 元田 浩 他「データマイニングの基礎」オーム社 (2006)
  機械学習手法の基礎とルールの検証
- J.Han and M.Kamber “Data Mining: Concepts and Techniques”second edition, Morgan Kaufmann (2006)
  多種多様な技法を網羅した本