公益財団法人
ひろしま産業振興機構

産×学 連携

3つのひかり 未来をつくる
広島市立大学
Hiroshima City University

Smart Factory推進Mgr養成 e-Learningコース

データマイニング概論

広島市立大学大学院
情報科学研究科
田村慶一

講義内容

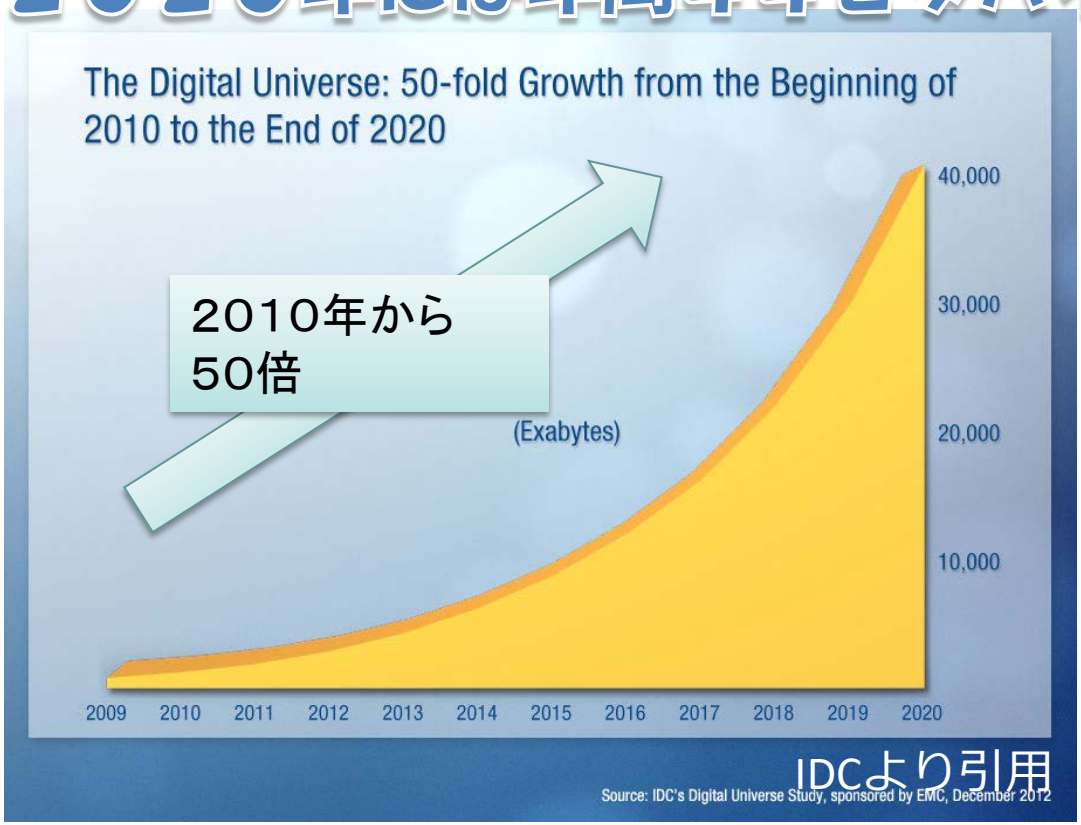
- データと知識発見
  - ビッグデータ，データ循環，データマイニングとは
- データマイニングの基礎技術
  - 決定木分析，クラスター分析，アソシエーション分析，主成分分析，回帰分析，異常検出
- 様々なメディアを対象としたデータマイニング
  - 時系列データマイニング，テキストマイニング，空間データマイニング

講義内容

- データと知識発見
  - ビッグデータ，知識発見，データマイニング，IoTとデータ循環
- データマイニングの基礎技術
  - 決定木分析，クラスター分析，アソシエーション分析，主成分分析，回帰分析，異常検出
- 様々なメディアを対象としたデータマイニング
  - 時系列データマイニング，テキストマイニング，空間データマイニング

ビッグデータ時代を迎えて

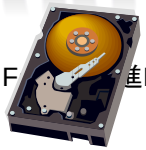
２０２０年には年間４４ゼタバイトが生成



単位

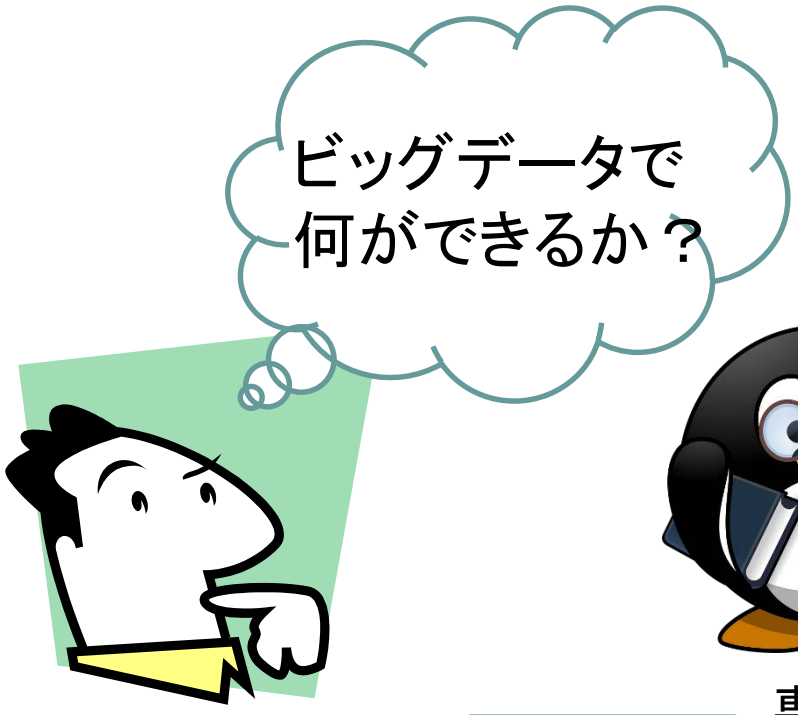
G（ギガ）10^9
T（テラ）10^{12}
P（ペタ）10^{15}
E（エクサ）10^{18}
Z（ゼタ）10^{21}

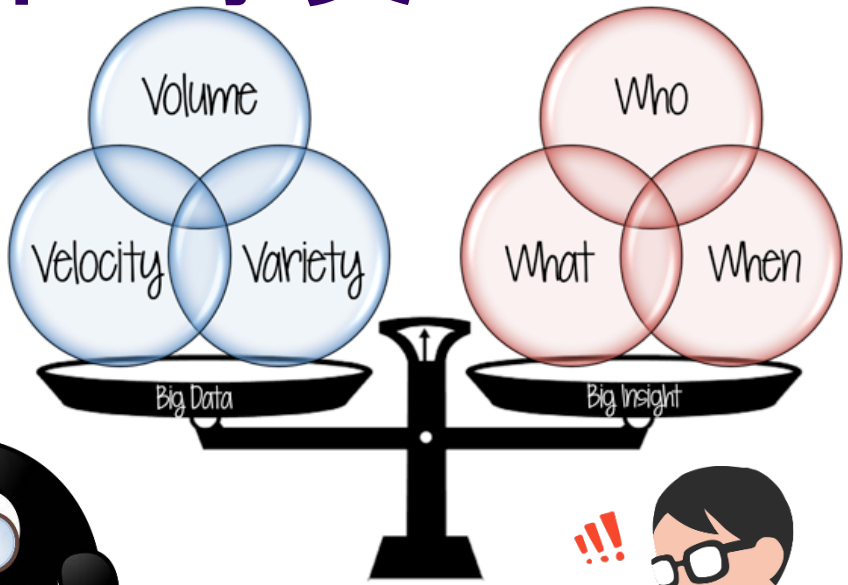
1,000,000,000,000,000,000,000



１テラのハードディスク

データ≒現実世界の事実





専門家

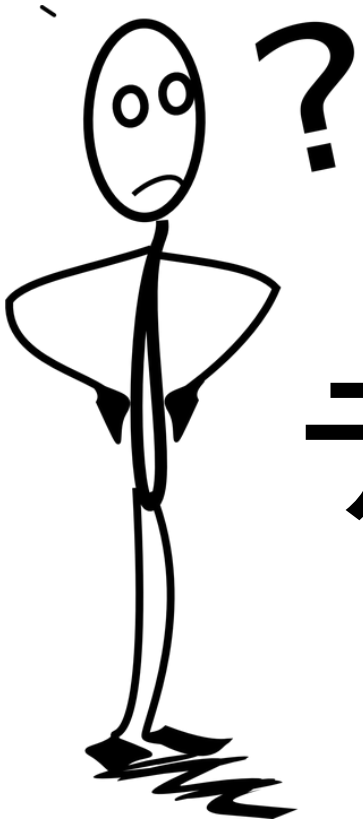
データサイエンティスト

経験
モデル
仮説

→ 法則
予測 ←

データの
分析

データの活用はできていますか



データ量＝知識量？

知識量 ≠ データ量

中世の人が一生で得るデータ量

＜　日刊紙のデータ量

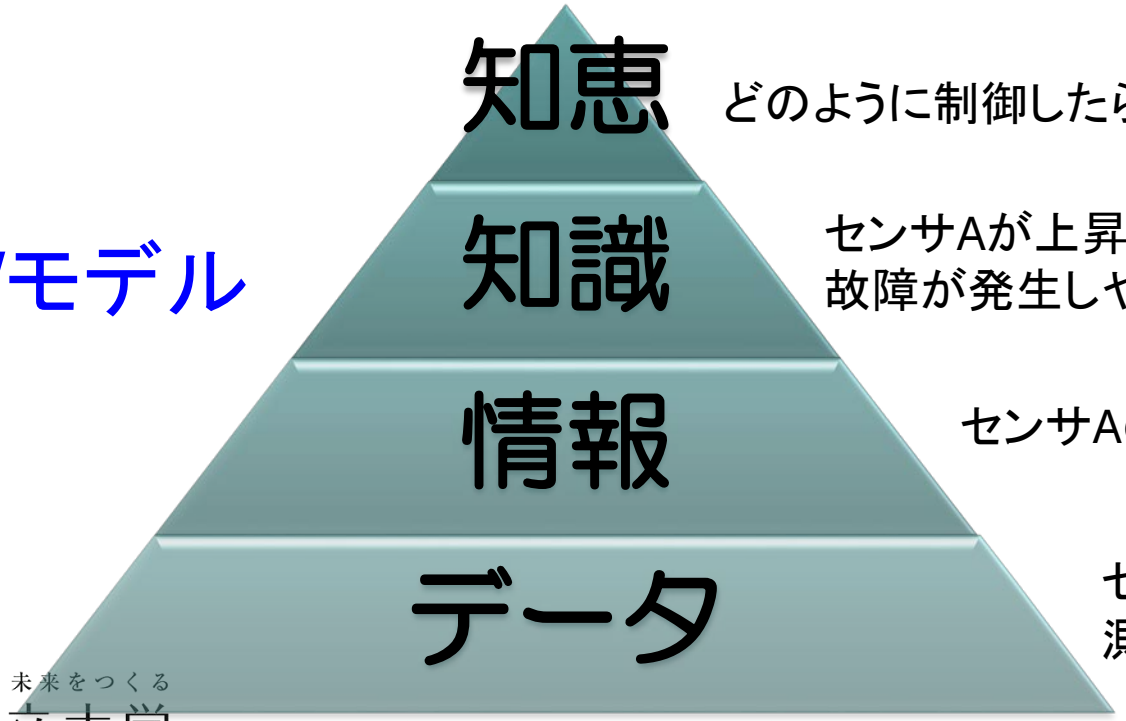
知識量は中世の人が多いのでは

数十Mバイト

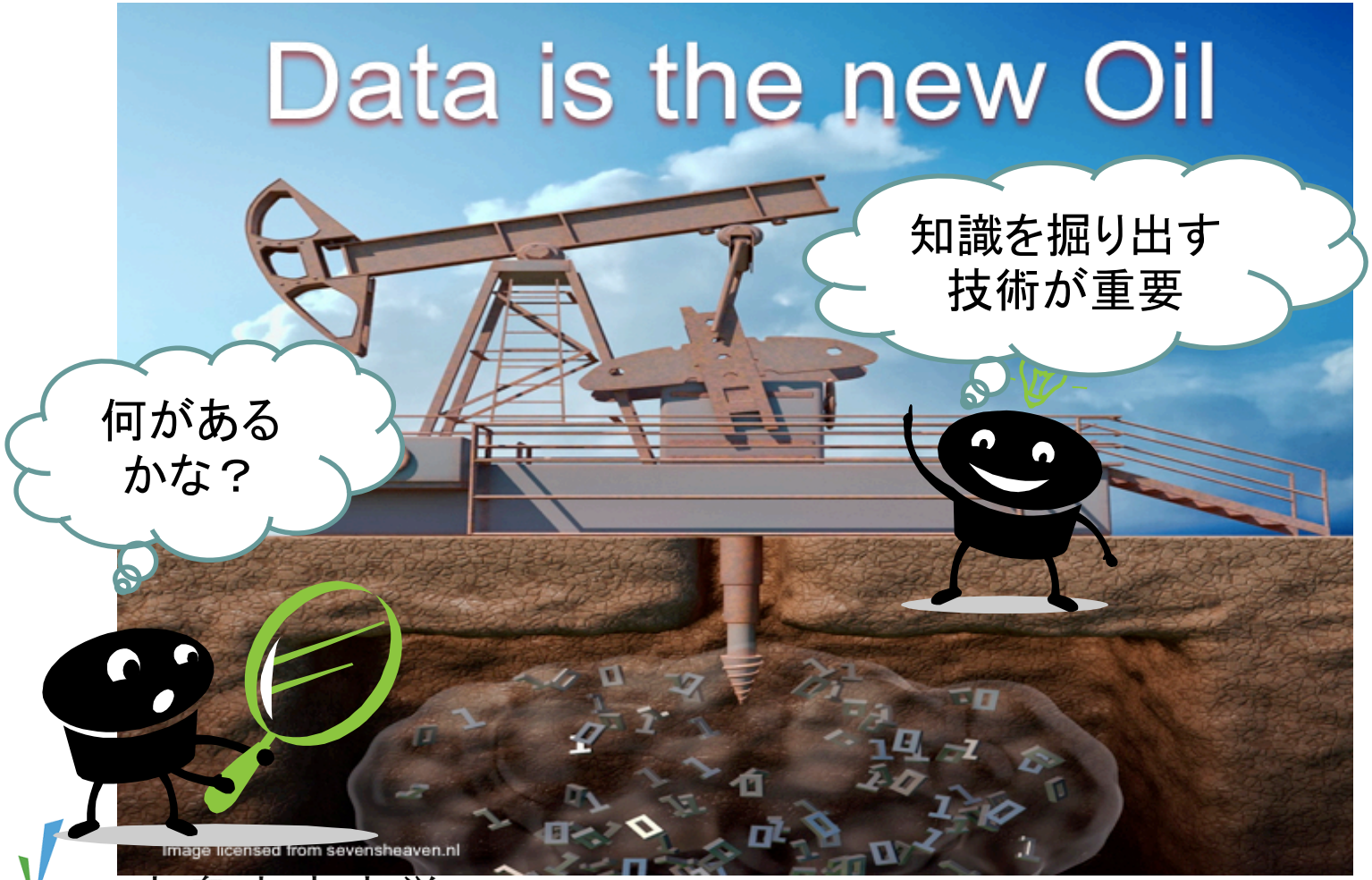
データ→情報→知識へ

- データから知識を取り出し，蓄積した知識を利活用する技術が重要

DIKWモデル



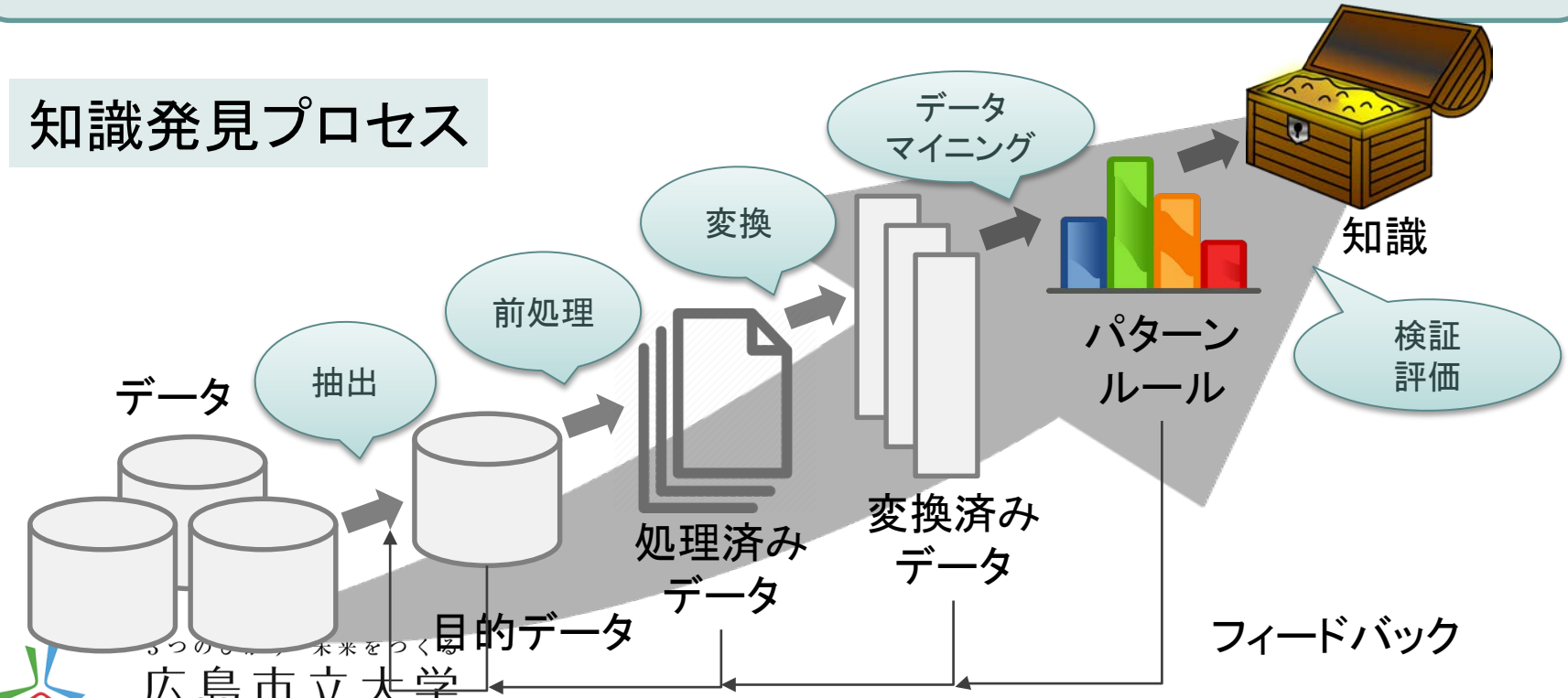
データは新しい資源



図はhttp://www.mediafuturist.comより引用

データマイニングとは

データの山の中から価値のある情報（パターンやルール）を見つけ出す技術

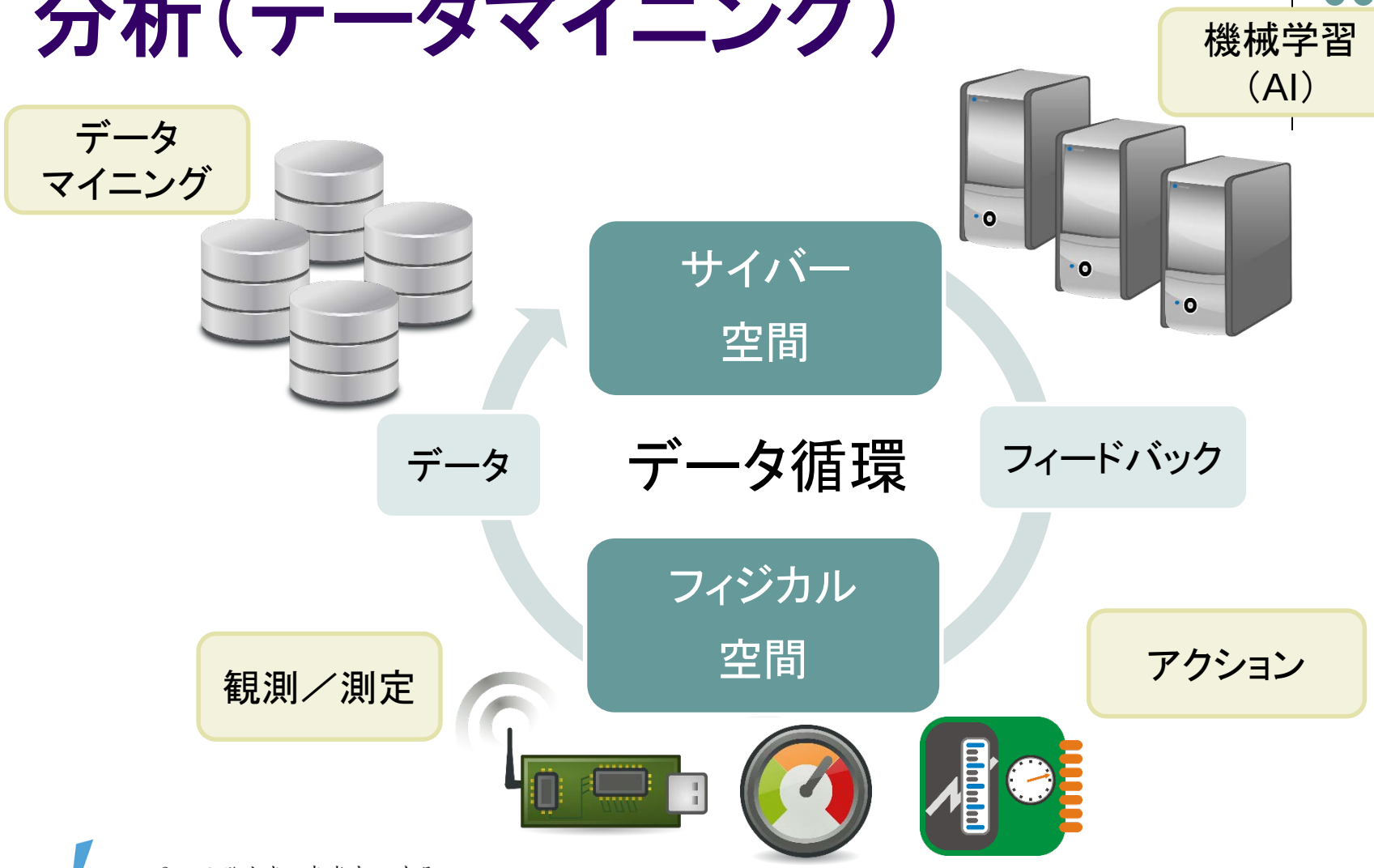


データマイニングの実例

- 飲料メーカ
  - 15本パックと7本パックは購入する顧客層が異なる→並べて販売すると両方の売り上げが増加
- レンタルビデオ店
  - 会員を「趣味別」及び「売上貢献別」に分類することで趣味に応じたクーポン発行やメール送付で売上向上
- ホームセンター
  - 売上データ，商品の陳列データ，従業員の行動データから顧客単価の高いスポットの特定し，店員の重点配置から売上げアップ

事例はhttps://liskul.com/wm_bd10-4861#1_12より引用
(2019年2月22日アクセス)

第4次産業革命＝IoT+AI+データ分析（データマイニング）



3つのひかり 未来をつくる

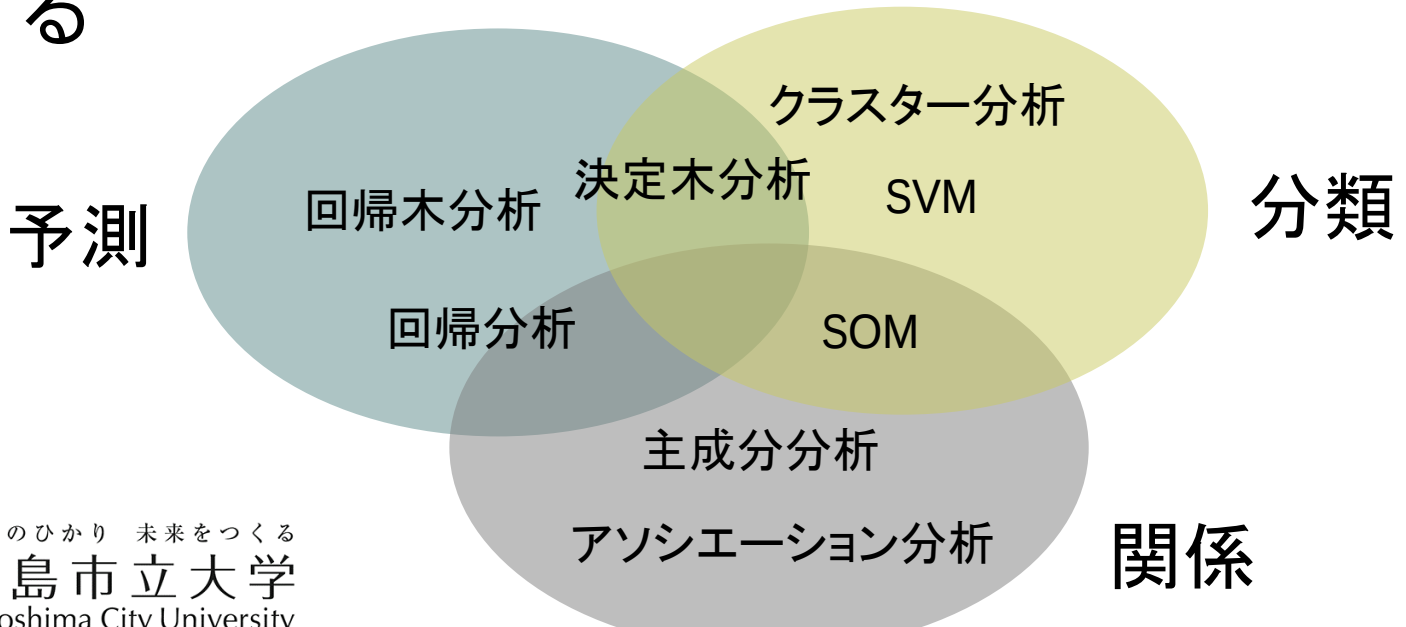
データマイニング＝データ循環を支える影の立役者

講義内容

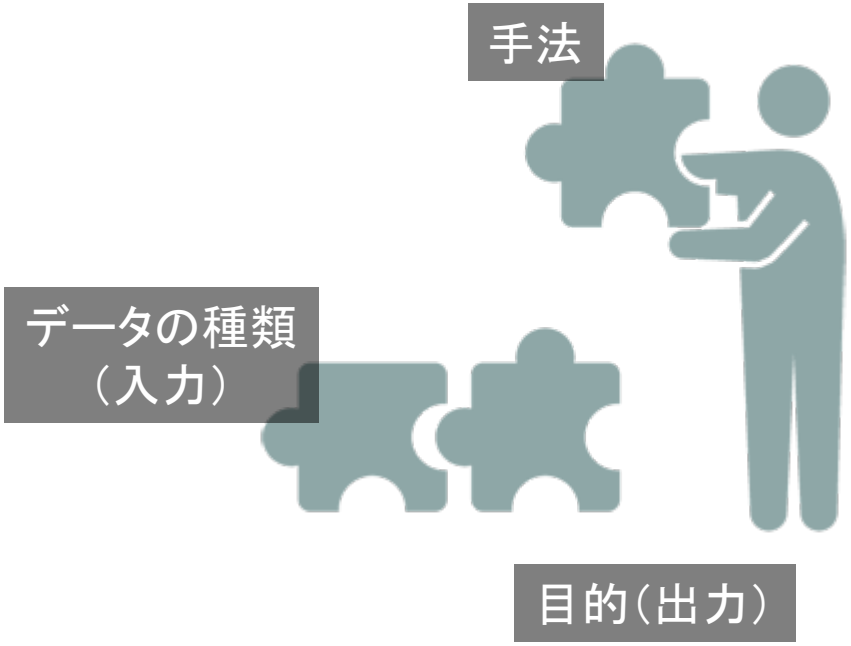
- データと知識発見
  - ビッグデータ，データ循環，データマイニングとは
- データマイニングの基礎技術
  - 決定木分析，クラスター分析，アソシエーション分析，主成分分析，回帰分析，異常検出
- 様々なメディアを対象としたデータマイニング
  - 時系列データマイニング，テキストマイニング，空間データマイニング

データマイニングの種類

- 予測：目的とする属性（変数）の値を予測する
- 分類：データを似たもの同士に分ける
- 関係：データ間や属性（変数）間の関係を明らかにする



データマイニングの基礎技術



データマイニングは魔法のランプ
ではない

目的にあった手法を
取捨選択する必要あり

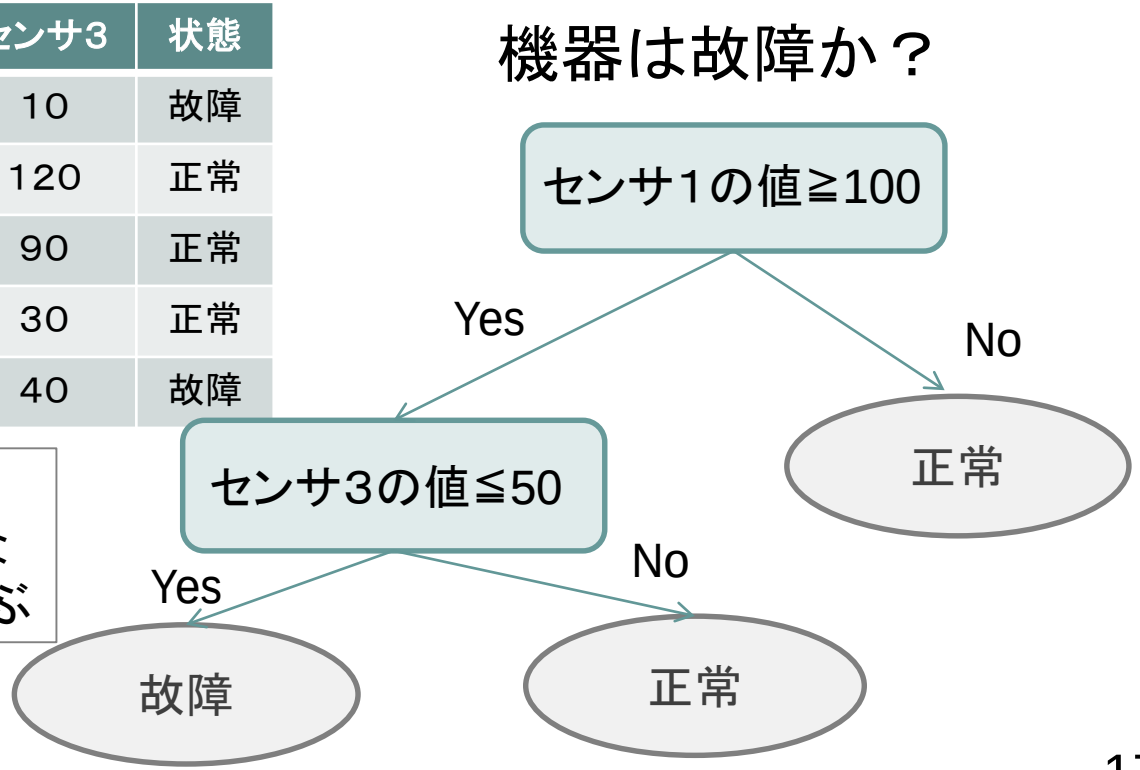
決定木分析

決定木（Decision Tree）

データの値
がXだったら
Yとなる

条件分岐の木構造を用いて分類を行うこと

	センサ1	センサ2	センサ3	状態
機器A	100	90	10	故障
機器B	110	50	120	正常
機器C	70	70	90	正常
機器D	30	20	30	正常
機器E	120	80	40	故障

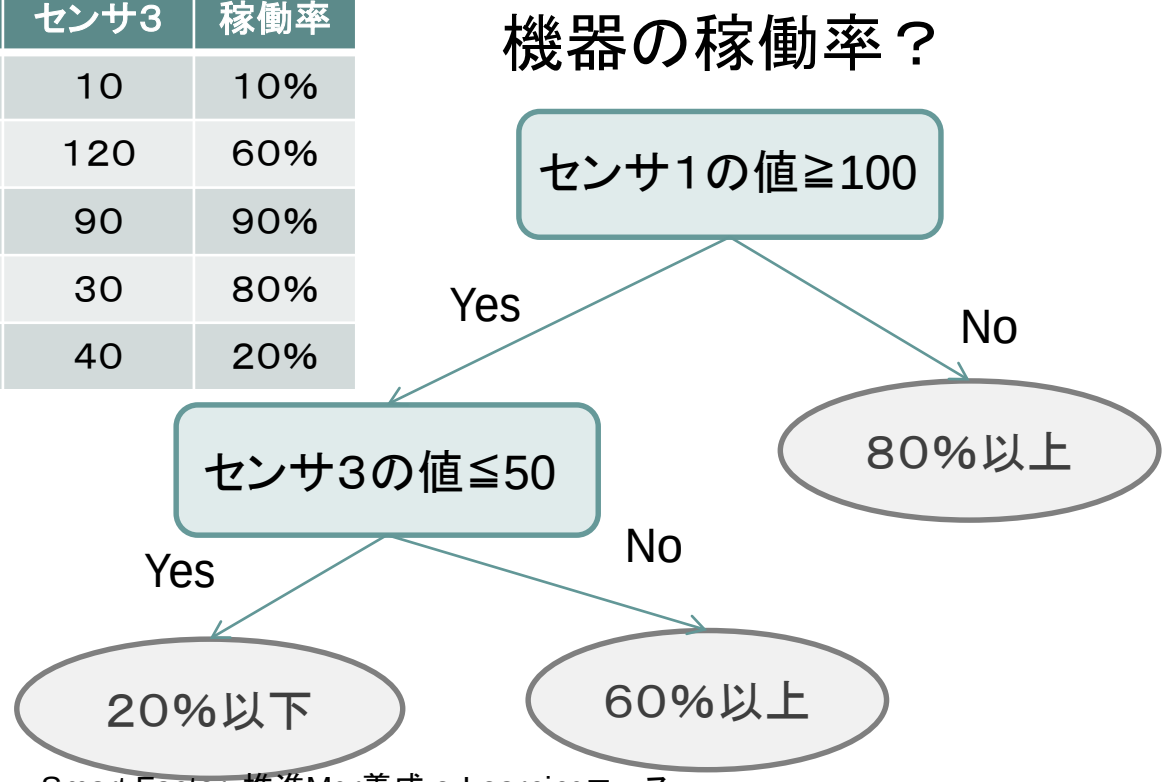


■決定木の作り方
不純度の差が大きくなるような
属性の条件分岐を優先的に選ぶ

回帰木（Regression Tree）

条件分岐の木構造を用いて回帰を行うこと

	センサ1	センサ2	センサ3	稼働率
機器A	100	90	10	10%
機器B	110	50	120	60%
機器C	70	70	90	90%
機器D	30	20	30	80%
機器E	120	80	40	20%



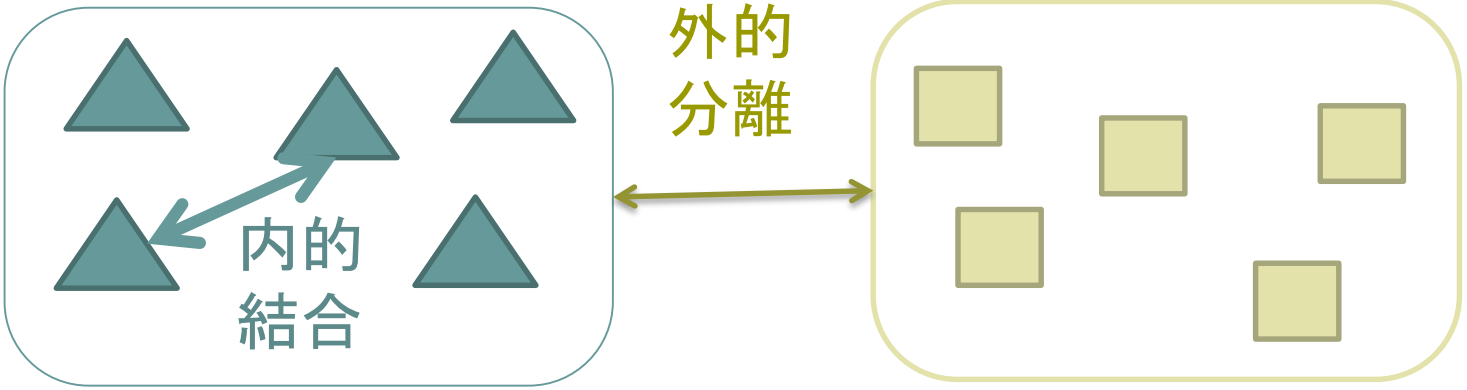
決定木／回帰木の活用シーン

- 小売業／サービス業全般
  - 対象データ：購入履歴データ，内容：商品購入層の把握（男性か女性か，また，10代か20代かなど）
- Eコマース全般
  - 対象データ：ユーザアクセスログ，内容：ユーザの趣向や購入動機の把握
- 製造業全般
  - 対象データ：計測データや稼働状況ログ，内容：不良品検出，故障予測

クラスター分析（クラスタリング）

クラスター（Cluster）分析

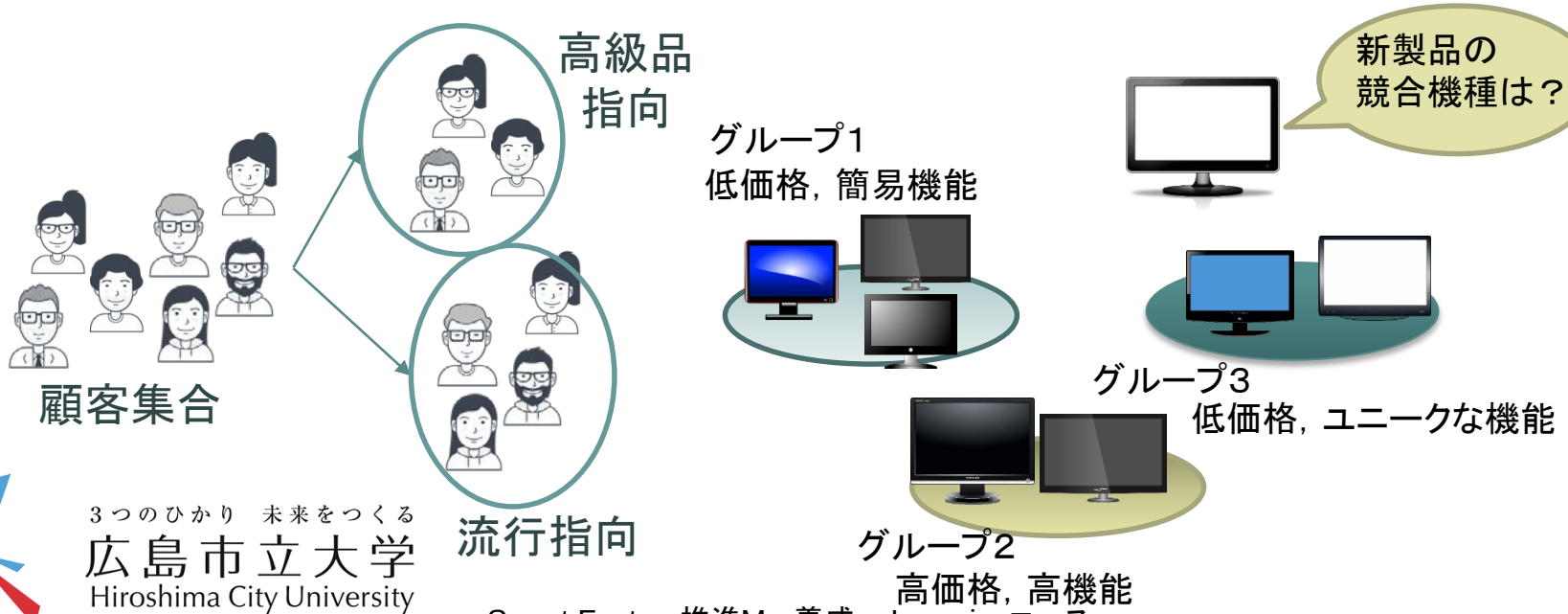
与えられたデータ集合をクラスタと呼ばれる
「まとまり」（部分集合）に分けること



まとまっているモノ　⇒　なんらかの性質や事象を示す
まとまっているデータ群　⇒　意味を持つデータ

クラスター分析の活用シーン

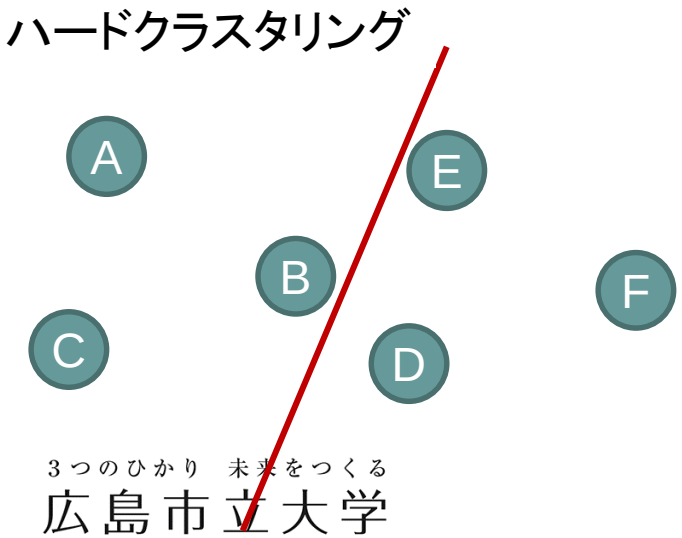
- 顧客／ユーザのグループ分け
- 製品／サービスのポジショニング
- 正常な機器と故障している機器の傾向把握
- 製品の品質管理

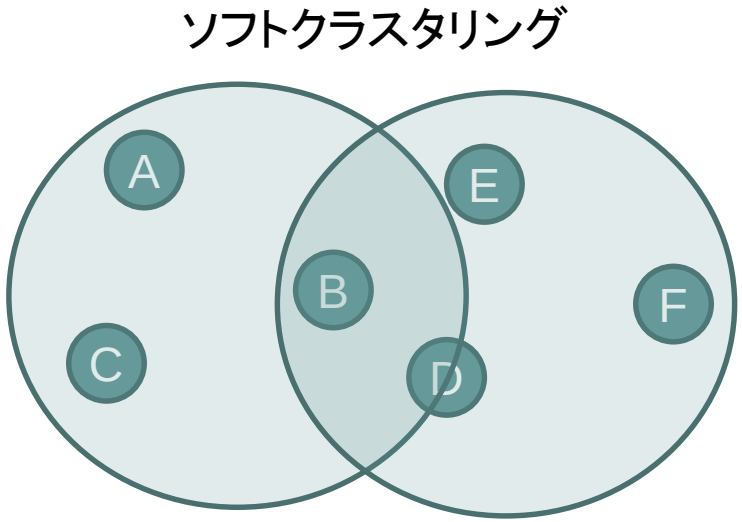


データクラスタリングの種類

クラスタリング＝クラスタに分けること

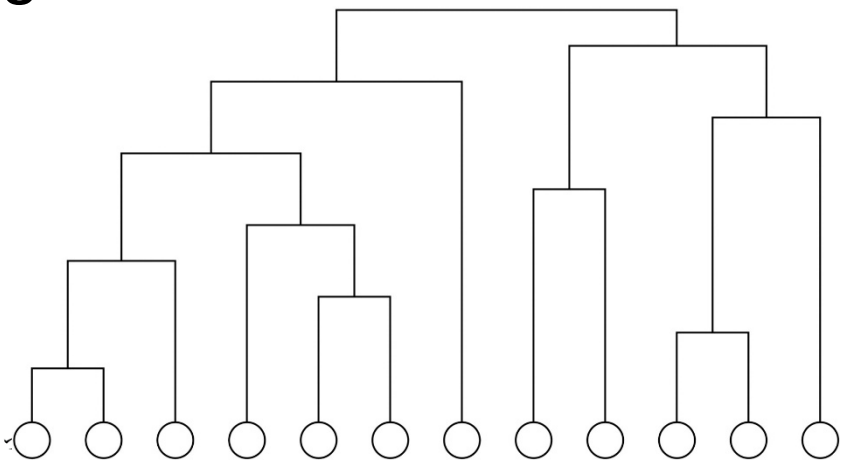
- **ハードクラスタリング**
  - データを別々のクラスタにきっちりと分ける
- ソフトクラスタリング
  - データは確率的に複数のクラスタに所属





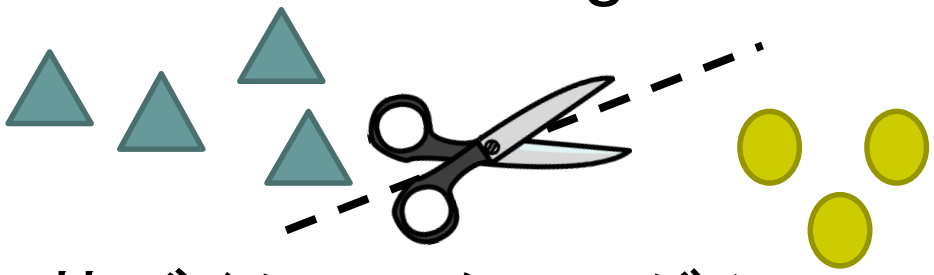
データクラスタリング手法（その1）

- 階層的クラスタリング（Hierarchical Clustering）
  - 凝集型階層的クラスタリング（Agglomerative Hierarchical Clustering）
  - 分割型階層的クラスタリング（Divisible Hierarchical Clustering）



データクラスタリング手法（その2）

- 分割最適化クラスタリング（Partition Optimization Clustering）



- 密度に基づくクラスタリング（Density-based Clustering）

その他　グラフ構造，SOMなど

アソシエーション分析

パターン

規則性は？

- データの中に存在する一定の規則や意味を持つ対象

データ中に現れる規則性＝注目すべきデータ，
現実世界のルールに結び付くデータ

アソシエーション（相関）分析

頻出するパターンを抽出し，その中から頻度の高い関係性（相関ルール）を抽出すること

購買履歴データ
{ポテトチップス，ガム，コーラ}
{お茶，団子，コーラ，パン}
{コーラ，ポテトチップス，パン}
{パン，ガム，飴}
{コーラ，ドーナツ，ポテトチップス，チョコレート}

ポテトチップスを買った人はよくコーラを買う

■相関ルールの抽出方法
アプリオリアルゴリズムを用いて無駄なパターンを見ないようにしている

マーケットバスケット分析 (Market Basket Analysis)

- 顧客はどの商品を組合せてよく買うか（一緒にどの商品を買うか）

POS（Point of Sales）データ

パン，ハム，牛乳
本，ガム，みかん

購買のパターンが知りたい

「おむつを買った人はビールを買う傾向がある」
1992年，Osco Drugsという小売ストア・チェーンの25店のPOSデータ

一般的な応用例（その1）

- 一緒に買われる商品は，近くに配置して販売

- 一緒に買われる商品はセットにして販売

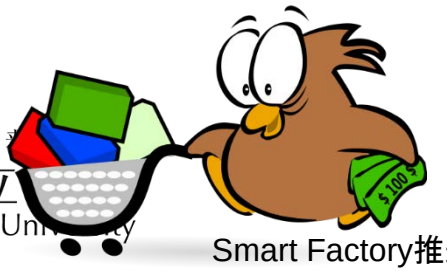
一般的な応用例（その2）

- 一緒に買われる商品は，片方をバーゲン価格に，もう一方は利益率を高くする

おにぎり100円！

おにぎりが割引！
お茶も買っていこう

- 一緒に買われる商品は，別々の場所に配置し，移動中に別の商品に目を向けさせる



おにぎりを買ったから
お茶はどこ？
あ！スイーツもいいな.

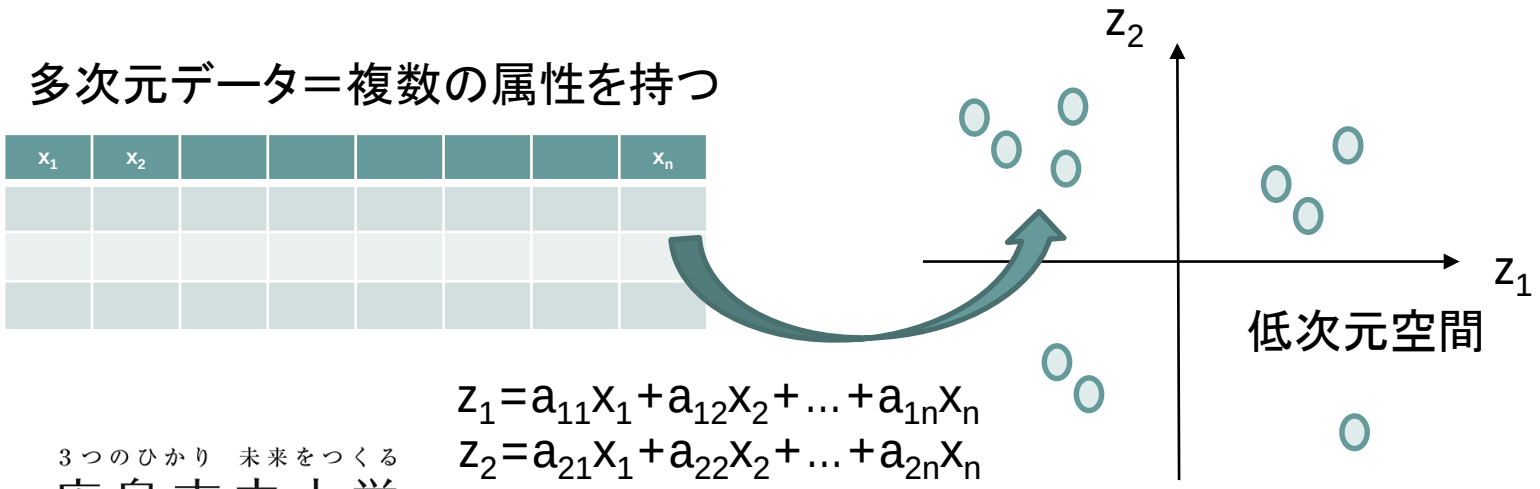
アソシエーション分析の活用シーン

- 環境データと，製造工程の中から，環境値と不良率の高い箇所のルールを取り出す
  - 例）天気が雨で◯◯を製造しているときに，△△という故障が発生
- 故障や欠陥の発生事例から，発生のパターンを取り出す
  - 例）製品の欠陥A，欠陥Bが現れるときに，欠陥Cがよく現れる

主成分分析

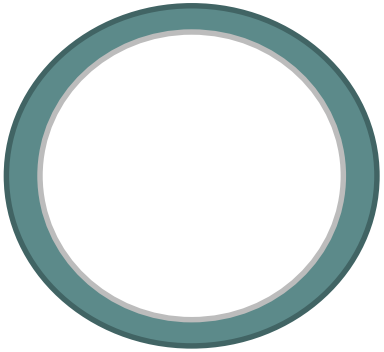
主成分分析 (Principal Component Analysis)

- 多次元データ(多変量データ)をより少数の主成分と呼ばれる指標に要約する分析手法
- 多次元データのもつ情報を損なわずに低次元空間に縮約する次元圧縮として利用される



主成分分析の活用シーン

- データマイニング手法の前処理
- 顧客満足度調査
- 製品のブランドイメージ調査



次元の呪い

高次元（属性が多い）だと
近いデータと遠いデータの距離の差がほとんどなくなる

↓

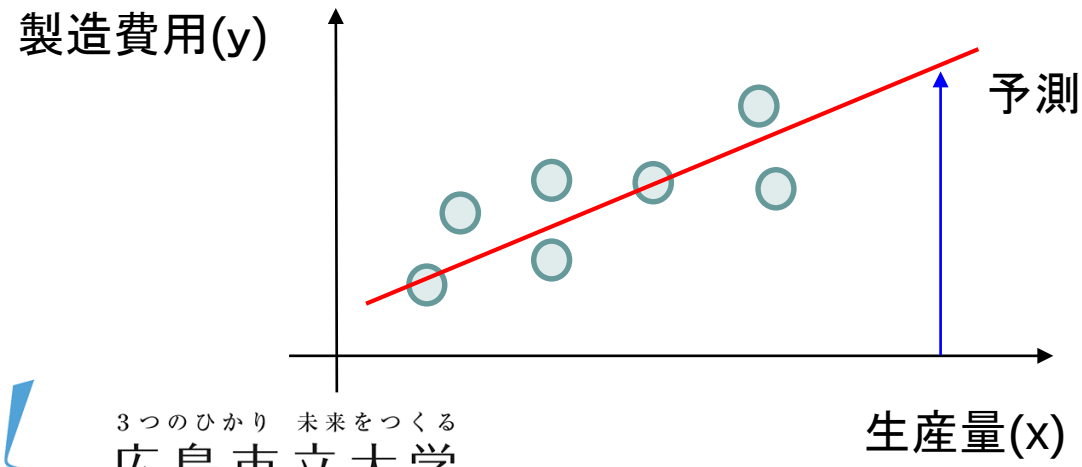
データマイニングの手法の多くがデータ間の距離を
ベースとしているため分析が難しくなる

回帰分析

回帰分析とは

因果関係があると思われる2つの変数について，一方の値（説明変数）で他方の値（目的変数）を表現

例）過去の生産量と製造費用のデータから将来の生産量に対する製造費用を予測



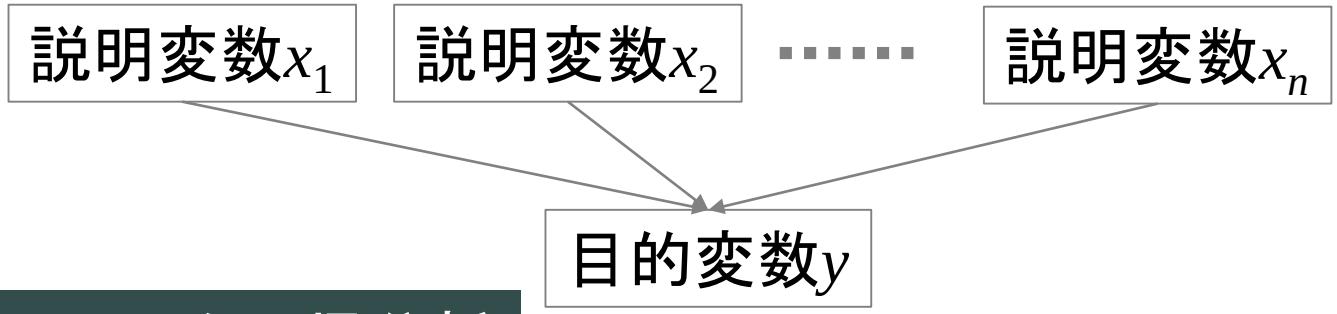
$y=ax+b$

過去のデータから係数を導き出す

重回帰／ロジスティック回帰分析

重回帰分析

複数の説明変数がある場合，複数の説明変数から
1つの目的変数を表現



ロジスティック回帰分析

目的変数が0か1のときの表現

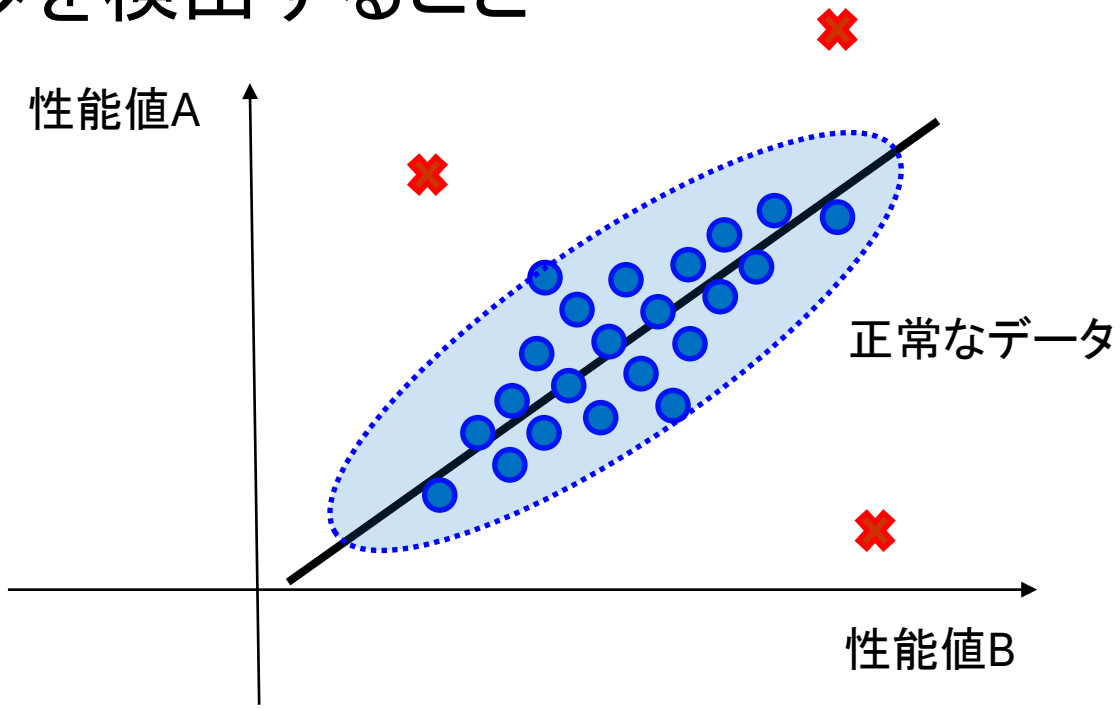
回帰分析の活用シーン

- 気象条件による販売量，欠陥率の予測
- 機器の稼働率から生産量を予測
- 欠陥率と商品購入の要因分析
- 試薬の量と副作用の可能性を予想
- 顧客層のスコアリング・要因分析

異常検出

異常検出（Anomaly Detection）

- 大多数のデータとは振る舞い（特徴）が異なるデータを検出すること

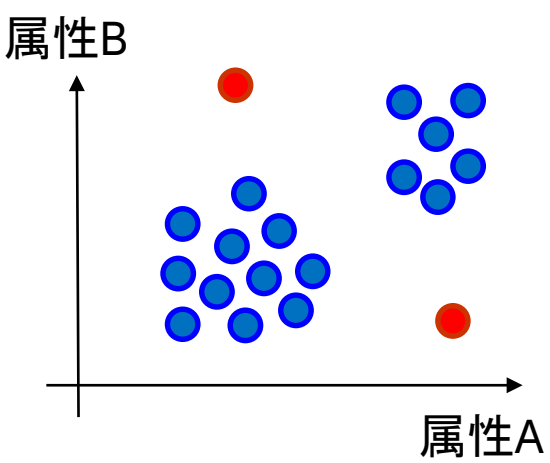


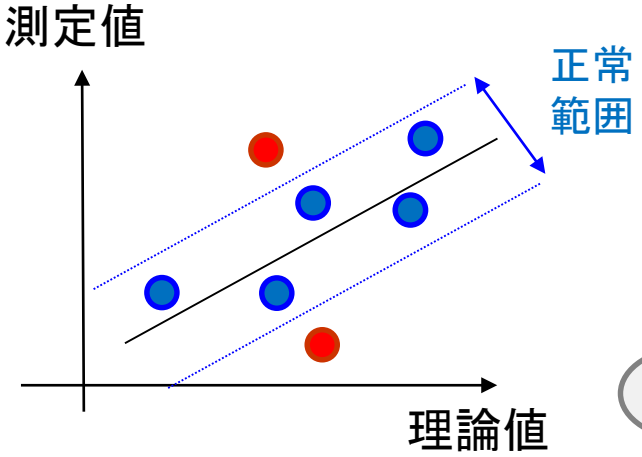
異常検出の活用シーン

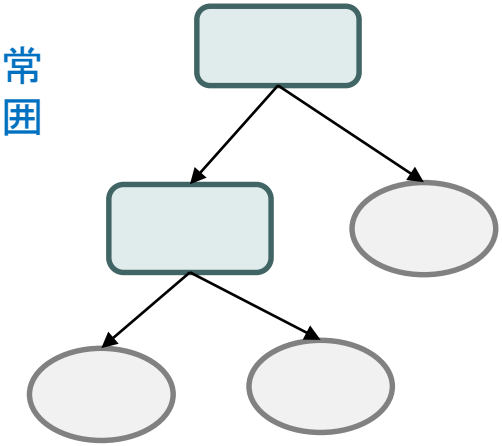
- 機械や設備の異常検知や予知保全
- 製品検査による不良品検出
- システム稼働／通信ネットワークの状態監視
- クレジットカードの不正利用検出
- 不正行動の検出

異常検出の手法

- 距離／密度に基づく検出手法
- 統計的分布に基づく検出手法
- ルールや分類に基づく検出手法







距離密度に基づく検出手法　統計的分布に基づく検出手法　ルールに基づく検出手法

講義内容

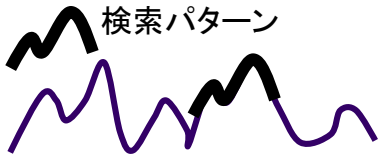
- データと知識発見
  - ビッグデータ，データ循環，データマイニングとは
- データマイニングの基礎技術
  - 決定木分析，クラスター分析，アソシエーション分析，主成分分析，回帰分析，異常検出
- 様々なメディアを対象としたデータマイニング
  - 時系列データマイニング，テキストマイニング，空間データマイニング

時系列データマイニング

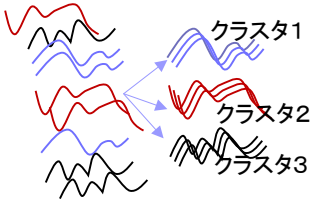
時系列データを対象としたデータマイニング

センサから集まるデータ
＝時系列データ

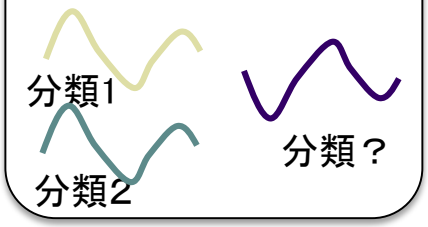
類似パターン検索



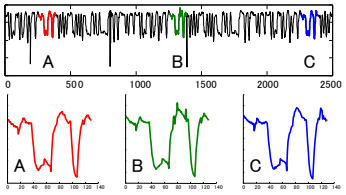
クラスタリング

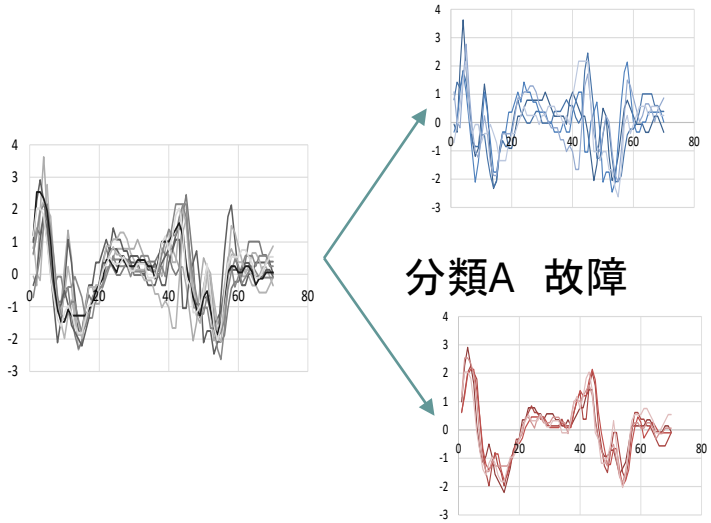


分類



モチーフ





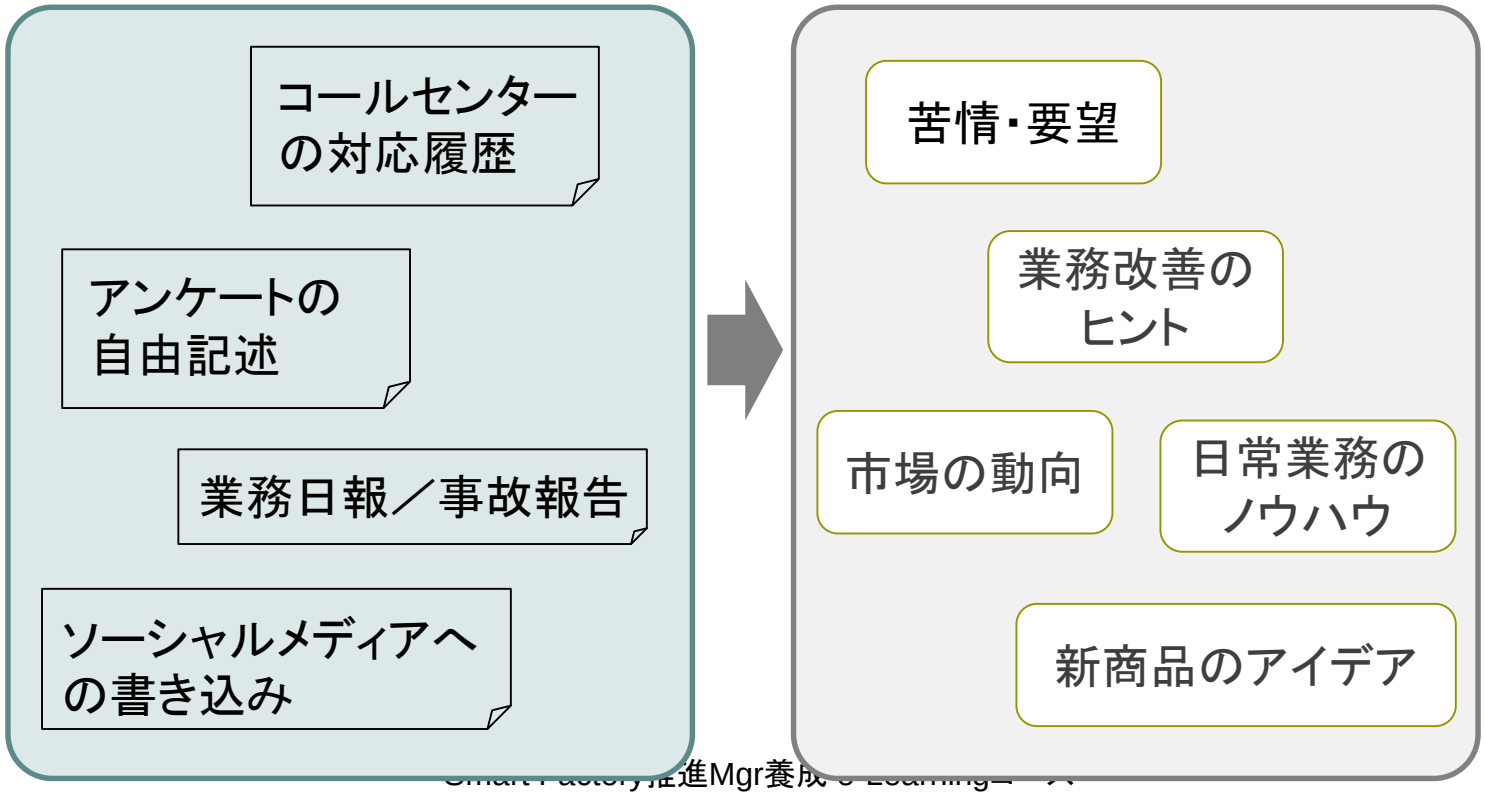
時系列データの分類問題

時系列データマイニングの活用シーン

- センサデータ全般のデータ分析
- 機器の状態判定
- 需要や生産量の予測
- 異常検出や故障判定

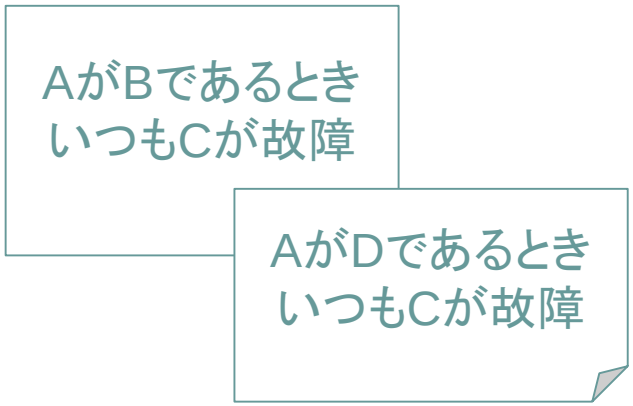
テキストマイニングとは

- 大量のテキストデータから出現頻度の高い語句，共起出現の相関などの有益な情報を取り出す



テキストマイニングの活用シーン

- コールセンターへの問い合わせ内容の分析
- 「顧客の声」の分析
- 業務日報の分析と暗黙知の形式知化
- ソーシャルメディアを活用したマーケティング



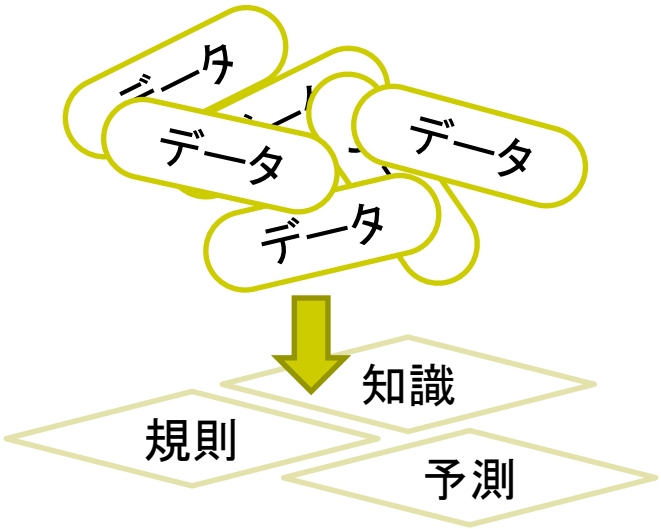
Aが{B, D}のときCが故障

頻出する語句や共起関係を分析

空間データマイニング

- 「空間データ」×「データマイニング」

データマイニング



空間データマイニング

いつ，どこで，何が，どのような状態か，どのようになるか

⇒　位置情報を他のデータと組み合わせて分析・活用する「ロケーション・インテリジェンス」を支える基本技術

空間データマイニングの活用シーン

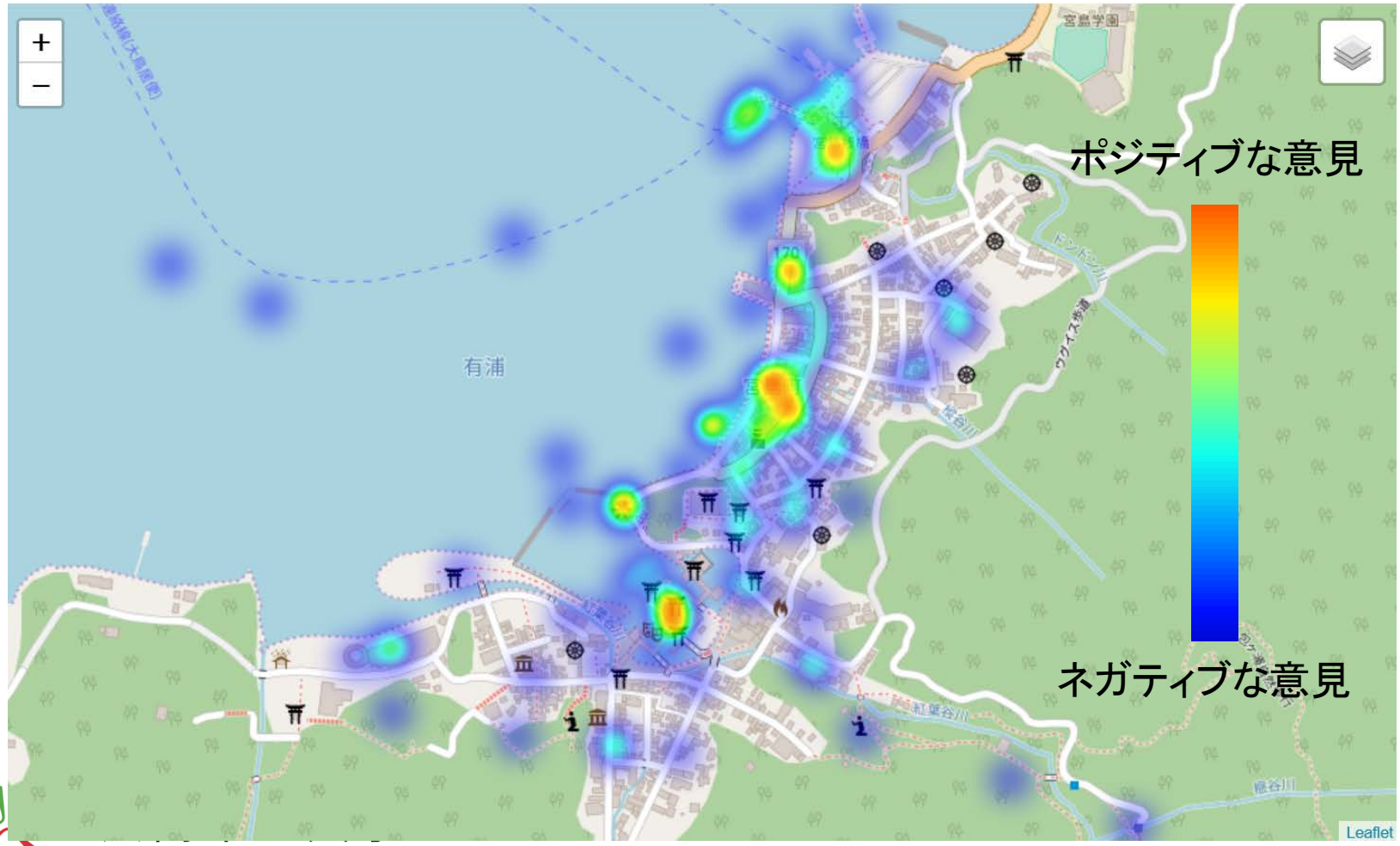
- 出店計画や商圏分析
- 顧客や従業員の行動分析
- 観光地の課題や魅力発見
- 配車システムの最適化
- 農地分析による生育状況，収穫量予測

動線管理

（出典：アプリックスのプレスリリース）

位置情報を利用した分析(ネガポジ)

広島市立大学観光関連データベース（WebAPI）から取得したツイートを分析



おわりに

おわりに

- データの中から価値のある情報（パターンやルール）を見つけ出す技術
- IoT成功のカギ，機器をつなげてデータを集めて表示だけでは不十分
- 目的やデータの内容に応じたデータマイニングの手法を選択する必要あり

小テスト

問題(1)

- 企業が保有する顧客や市場などの膨大なデータから，有用な情報や関係を見つけ出す手法はどれか(基本情報技術者 平成24年春期 午前問64)。

ア　データウェアハウス

イ　データディクショナリ

ウ　データフローダイアグラム

エ　データマイニング

問題(2)

- データマイニングの事例として，適切なものはどれか。

ア　あるセンサの測定値の最大値を調べた

イ　従業員の出勤時間の平均をグラフで描いた

ウ　あるセンサの故障率と室温の関係を調べた

エ　あるセンサの過去3年の故障率を調べた

問題(3)

- 観測されたデータについてある属性値を使って他の属性値を表現する式を求めるデータマイニング手法はどれか。

ア　クラスター分析

イ　テキストマイニング

ウ　回帰分析

エ　主成分分析

問題（4）

- 観測されたデータ集合について類似したデータのグループに分割するデータマイニングの手法はどれか。

ア　クラスター分析

イ　テキストマイニング

ウ　回帰分析

エ　主成分分析

問題（5）

- 複数の属性を持ち多次元のデータを低次元のデータに次元圧縮するデータマイニングの手法はどれか。

ア　クラスター分析

イ　テキストマイニング

ウ　回帰分析

エ　主成分分析

小テストの解答

- 問題（1） エ
- 問題（2） ウ
- 問題（3） ウ
- 問題（4） ア
- 問題（5） エ

教科書

- 豊田 秀樹，データマイニング入門，東京図書，2008年
- 元田 浩ほか，データマイニングの基礎 (IT Text) ，オーム社，2006年
- 福田 剛志ほか，データマイニング (データサイエンス・シリーズ 3)，共立出版，2001年
- Foster Provostほか，戦略的データサイエンス入門 ―ビジネスに活かすコンセプトとテクニック，オライリージャパン，2014年