Human in the Loop

情報科学概論 1

高間　康史　（3回目講義）

データサイエンスとは

- データから価値を引き出す
- BI (Business Intelligence)
  - 企業内外に蓄積されたデータを組織的・系統的に集約・整理・分析
  - ビジネス上の各種意思決定に有用な知識・洞察を生み出す
  - 情報部門 ➡ エンドユーザ（経営者・一般社員）
- BIとの違い[城田12]：データのライフサイクル全体に関与
  - OJTでは不十分：教育・学習の必要性
- AI戦略2019（統合イノベーション戦略推進会議）
  - 文理を問わず，一定規模の大学・高専生が自らの専門分野への数理・データサイエンス・AIの応用基礎力を習得すること

大阪ガス全社員「データ分析官」 【日経新聞2020.12.17】
電力部門含め5000人が検索・共有、営業力向上で競争勝ち抜く

- 大阪ガスが社内のデータ活用の仕組みを変える．あらゆる社員がガス供給量や機器稼働状況などのリアルタイムデータを検索できるようにし，商業施設の省エネ提案や家庭用燃料電池「エネファーム」の販売などで攻勢をかける．全社員を「データ分析官」に育て，業界をまたいだ競争で生き残る．
- ・・・
- 原動力は，11年度から進めてきたデータサイエンス研修だ．分析を進めるプロセスや注意点などを教え込み，19年度までに累計受講者は1900人に達した．実態を熟知する現場社員がスキルを習得することで，分析担当に頼り切らず新たな戦略を自発的に立案しやすくなる．・・・

データサイエンティストに必要なスキル

計算機科学

データ工学

AI

数学・統計

ドメイン知識

データ分析の分類

- 検証的データ分析
  (Confirmatory Data Analysis, CDA)
  - あらかじめ持っている仮説の検証
  - 統計的検定
  - データは後から集める（実験条件の設定）
- 探索的データ分析
  (Exploratory Data Analysis, EDA)
  - データを眺めながら仮説を生成
  - すでにあるデータが出発点
  - データマイニング，可視化

データマイニングとは

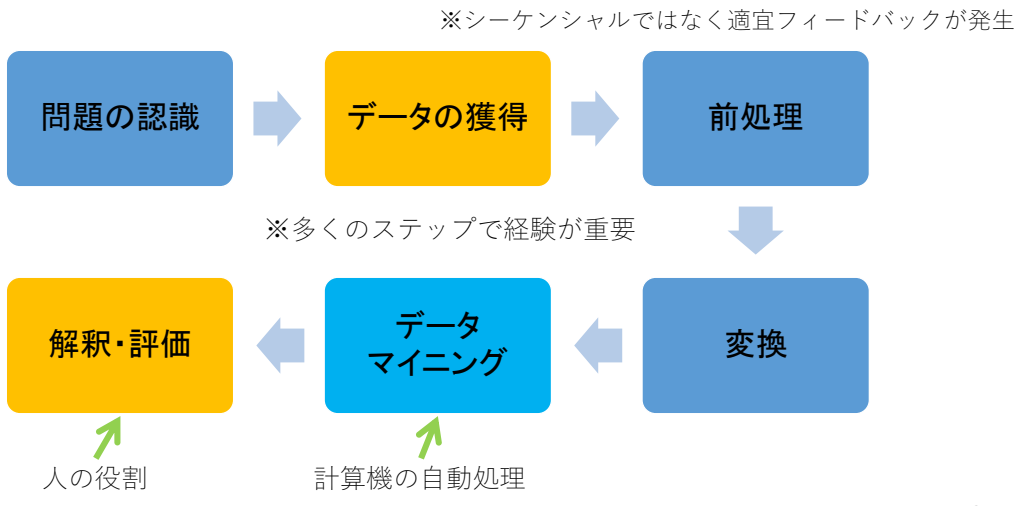
- 『勘』から『根拠のある判断』へ
  - ベテランの経験による判断：暗黙知
  - 伝承，未知事象への対応が困難

- 大規模データの活用
  - インターネット，センサの進歩・普及 ➡ 取得が容易に
  - 記憶装置の大規模化・クラウドの普及 ➡ 蓄積が容易に
  - データに内在する構造・規則性の発見

データマイニングに対する誤解

- 専門家の経験・勘による判断は間違い？
  - 専門家：時間をかけて，暗黙知的に獲得
    ⇐　未知事象・変化の激しい状況に対応困難
  - 計算機：高速，明示的に獲得
- 計算機に任せておけばよい？
  - 知識発見プロセスの一部の自動化に過ぎない
  - 人間と計算機の役割分担が重要
  - 最終判断・意思決定は人間の役割

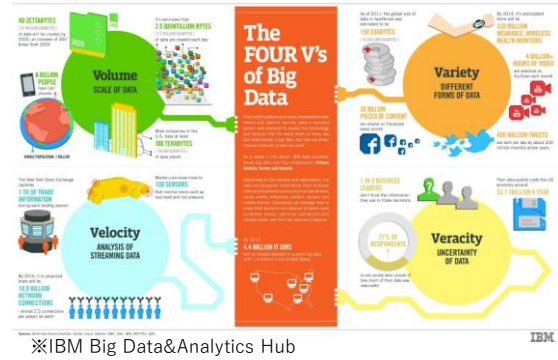
知識発見プロセス
KDD: Knowledge-Discovery in Database



データの獲得

- すでにデータは存在している場合が多い
  - 「とりあえず貯めておく」ことが可能
  - いろんな部門から集めてくる
  - 不足分のみ収集
- 色々なデータに様々な価値
  - 購買・閲覧記録
  - アンケートの自由回答
  - 日報，報告書
  - 各種センサデータ
  - レビュー（口コミ）
  - 位置情報

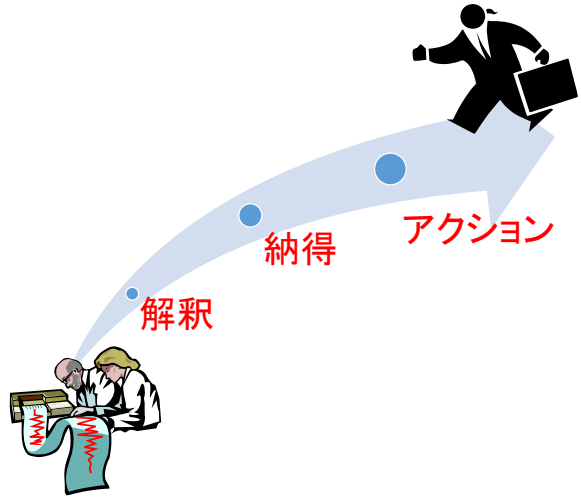
規模（Volume）
多様性（Variety）
速度（Velocity）
不確かさ（Veracity）



※IBM Big Data&Analytics Hub

解釈と評価

- 獲得した知識をアクションへ
- 解釈のしやすさ
  - 表現形式
  - シンプル・本質的
- 納得する『根拠』
  - 精度（定量的）
  - シナリオ・意味づけ（定性・主観的）

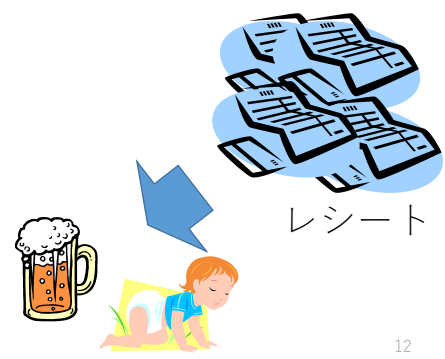


データマイニングの代表的手法

- 興味深いパターンを発見したい
  - 相関ルール（association rule mining）
- データ空間の大まかな構造を知りたい
  - クラスタリング
- 将来（未知データ）の予測をしたい／判断を自動化したい
  - 分類モデル構築
  - 回帰分析

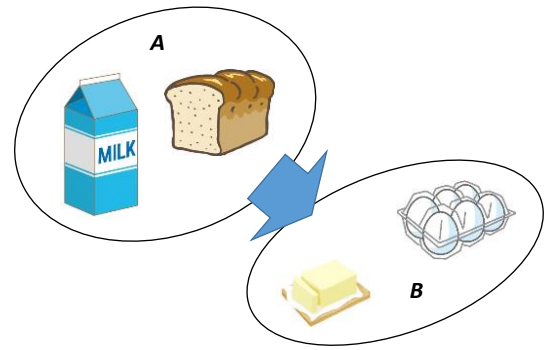
相関ルール
Association Rule

- アイテムの共起パターンを知る
- 応用例
  - バスケット分析：同時に購入される商品の分析
  - 行動パターン（動作系列）の分析
    - 時系列相関ルールマイニング
- 興味深いパターンの発見
  - 興味深さの指標とは？



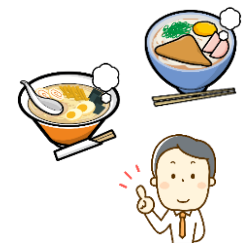
相関ルールの定義

- I… アイテム集合
- 相関ルール：　$A \Rightarrow B$
  - A に含まれるアイテムを購入した人は，B に含まれるアイテムも一緒に購入
  - $A, B \subseteq I$, $A \cap B = \varnothing$
- 例
  - I = {牛乳，パン，バター，卵}
  - {牛乳，パン} => {バター，卵}



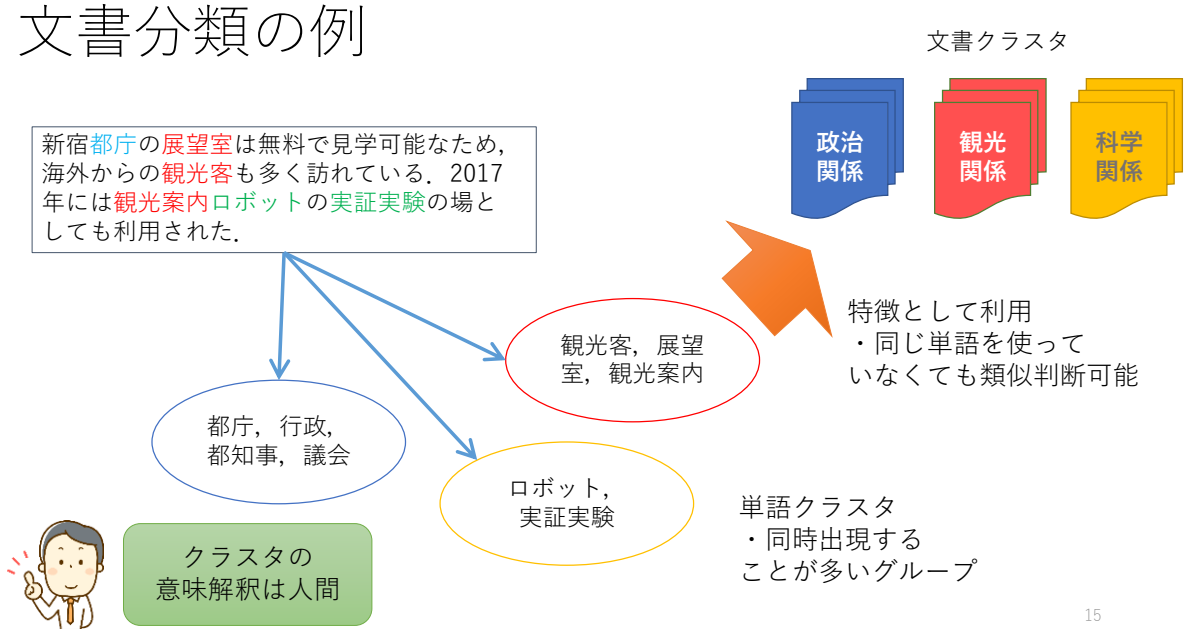
クラスタリング

- データ集合を，類似するいくつかのグループに分割
  - グループ＝クラスタ
  - 顧客のタイプ分け（cf. セグメンテーション）
  - 故障原因・不具合の分類
  - 競合商品の分析: 自社製品と比較される他社製品は？
- セグメンテーション（広告業界）
  - F1層：20～34歳女性
  - M2層：35～49歳男性
  - T層：男女13～19歳



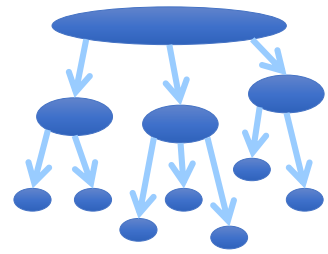
データに内在する
グループの発見

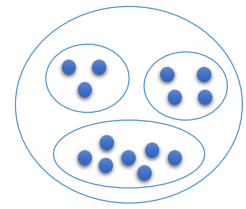
文書分類の例



クラスタリング手法の分類

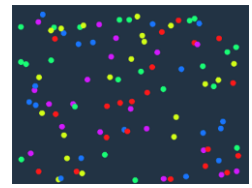
- 分割　vs. 併合
  - 分割：トップダウン
    全体集合→小さいグループ
  - 併合：ボトムアップ
    個々のデータ→全体集合
- 階層的 vs. パーティション（partition）
  - 階層的：結果が階層構造（樹形図）
    - クラスタ間関係：上下関係，兄弟関係
  - パーティション：フラットに分割

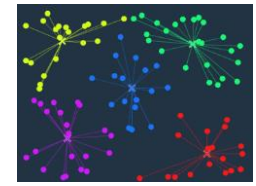




K-means

- 代表的クラスタリング手法
- 特徴
  - 分割・パーティション
  - クラスタ数はユーザが指定
  - 高速
  - シンプル
    - 実装，結果の理解が容易
    - 拡張が容易
  - △初期値依存

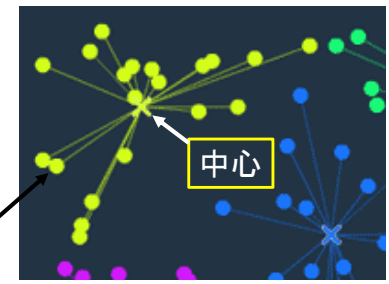




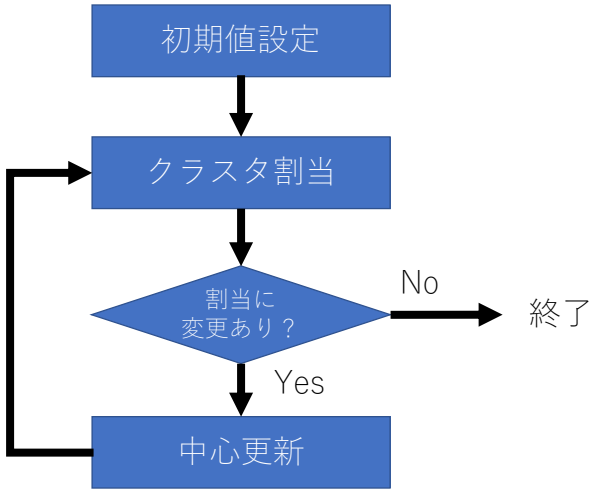
アルゴリズムの説明 (1)

- 入力データ
  - 距離行列 or 類似度行列：データ間の距離（類似度）を定義
  - 各データのベクトルが与えられ，そこから計算する場合も
    - 距離：ユークリッド距離，類似度：余弦，内積
  - クラスタ数：K
- 出力（クラスタリング結果）
  - クラスタ中心：重心ベクトル
  - クラスタメンバ：属するデータ

} K個出力



アルゴリズムの説明 (2)

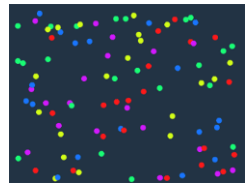


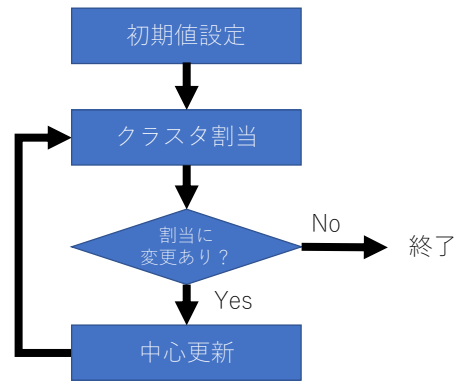
※初期値設定方法によっては
クラスタ割当の前に中心更新
をする場合もあり

アルゴリズムの説明 (3)

- *K*個のクラスタ中心をランダムに決定
  - 乱数でベクトルを生成
  - データから*K*個ランダムに選択
- 全データを*K*個のクラスタにランダムに割当
  - クラスタ割当前に中心更新を行う必要

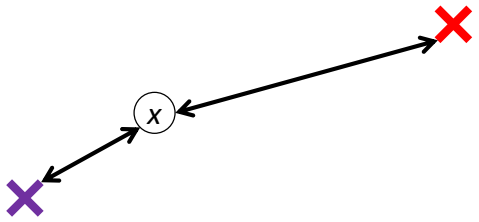
※色＝クラスタ

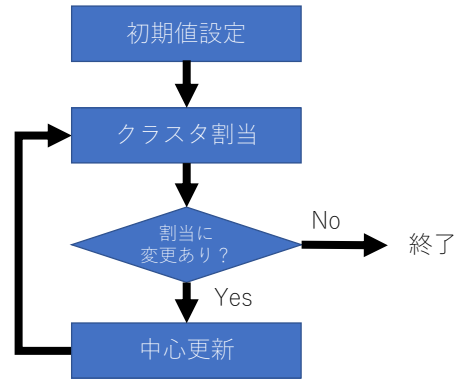




アルゴリズムの説明 (4)

- 各データ $\boldsymbol{x}$ について
  - $\boldsymbol{x}$ に一番近いクラスタ中心 $\boldsymbol{p_c}$ を探す
  - $\boldsymbol{p_c}$ に対応したクラスタ $\boldsymbol{C}$ に割り当てる
- 全データについて実施



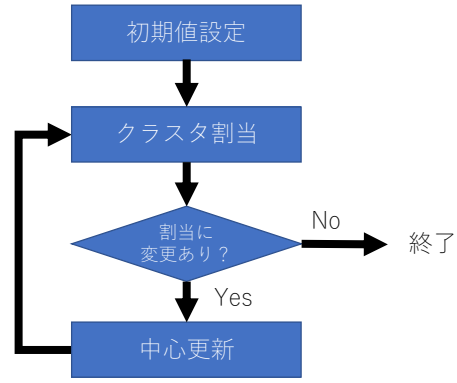


アルゴリズムの説明 (5)

- クラスタ中心 $\boldsymbol{c}$ の更新

$$p_c = \frac{1}{|C|}\sum_{x \in C} x$$

- 全クラスタについて更新

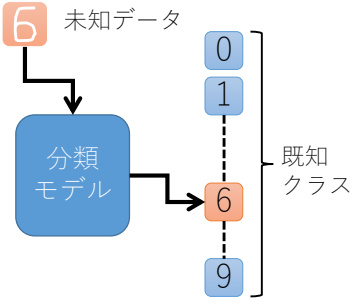


分類モデル構築とは

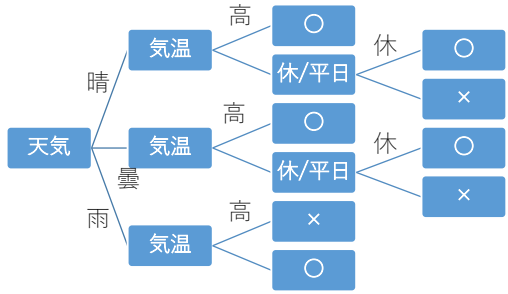
- 未知のデータを，既知のクラスに分類
  - グループ分けという点では，クラスタリングと同様
- クラスタリングとの違い
  - **教師あり学習**：訓練データから学習
    - グループ（クラス）の個数・意味が既知
    - 訓練データ（所属クラスが既知）から**分類モデル構築**
  - **教師なし学習**：訓練データなし　←　クラスタリング
    - グループ（クラスター）の個数・意味が未知
    - グループの意味は後から考える

教師あり学習の例

- パターン認識
  - 数字認識，文字認識
  - 指紋認証，顔認証：本人か否か
  - 顔認識（デジカメ）：顔か否か
- フィルタリング
  - スパムメールフィルタ
- マーケティング
  - 継続顧客／離反顧客の分類
  - 仕入れ予測：売上の良い日／悪い日の分類
  - 与信：返済の可能性を予測



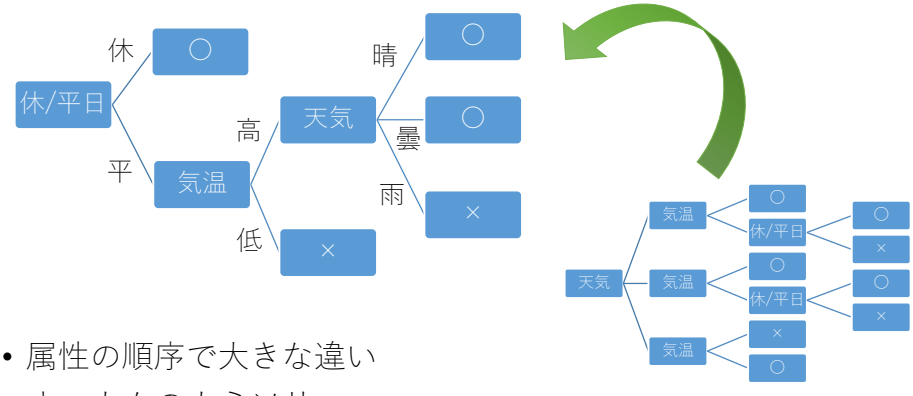
代表的分類モデル構築手法
決定木（Decision tree）



天気	気温	休/平日	売上
曇	高	休日	○
晴	低	休日	○
雨	低	休日	○
曇	高	平日	○
晴	高	平日	○
曇	低	休日	○
雨	高	平日	×
曇	低	平日	×
晴	低	平日	×
曇	低	平日	×

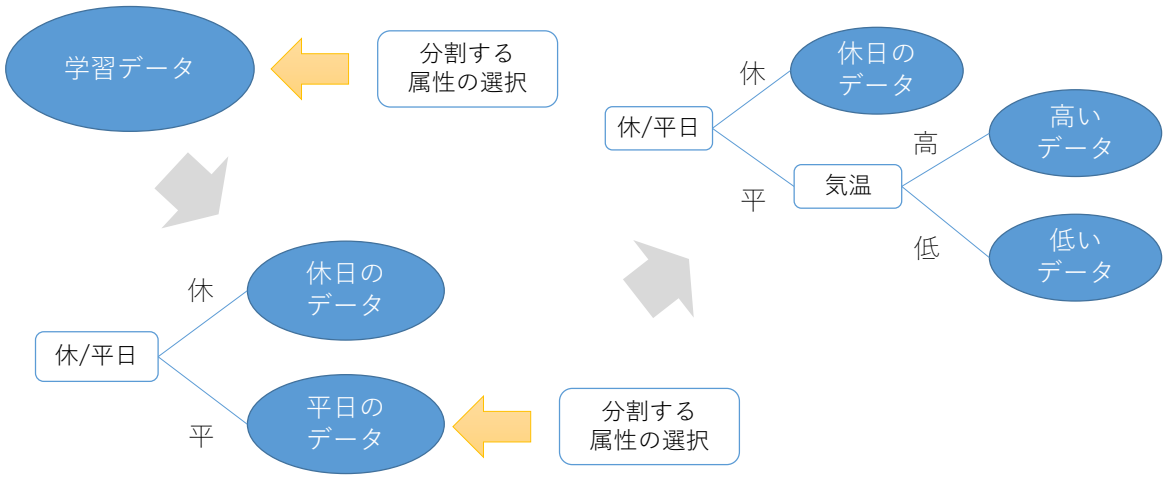
- ノード＝属性，エッジ＝属性値
- パス（ルート→リーフ）
  ＝分類規則
- 可読性が高い

良い決定木とは



- 属性の順序で大きな違い
- オッカムのカミソリ
  - 現状を同程度説明する仮説なら，よりシンプルなものを選ぶべき

アルゴリズムの概要



属性の選択基準：不純度

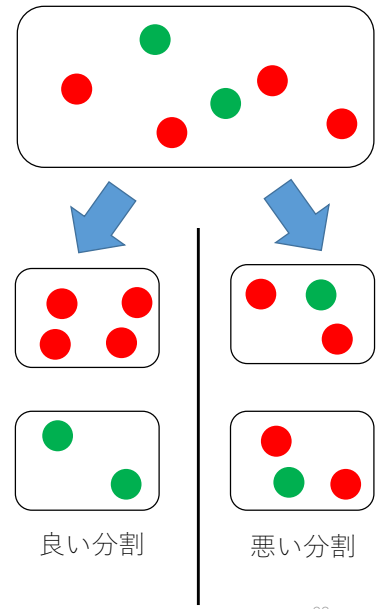
- 分割における多様性の指標
  - 分割により多様性が減少することが望ましい
  - 多様性の減少：同一クラスのデータが集合
- 代表的指標：大きいほど不純
  - エントロピー（平均情報量）

$$Ent = -\sum_{c=1}^{J} p_c \log_2 p_c$$

  - Gini分散指標

$$Gini = 1 - \sum_{c=1}^{J} {p_c}^2$$

J: クラス数
p_c: cクラスの割合



例題：最初にどの属性を選ぶか？

- 属性：天気
  - 曇：5件．〇:3，×:2
    - Ent = -0.6Log0.6-0.4Log0.4=0.971
  - 晴：3件．〇:2，×:1
    - Ent = 0.918
  - 雨：2件．〇:1，×:1
    - Ent = 1
  - 加重平均：0.961
- 属性：気温
  - 高：4件．〇:3，×:1
    - Ent = 0.811
  - 低：6件．〇:3，×3
    - Ent = 1
  - 加重平均：0.924
- 属性：休／平日
  - 休日：4件．〇:4，×:0
    - Ent = 0
  - 平日：6件．〇:2，×:4
    - Ent = 0.918
  - 加重平均：0.551

◎休／平日で分割

天気	気温	休/平日	売上
曇	高	休日	〇
晴	低	休日	〇
雨	低	休日	〇
曇	高	平日	〇
晴	高	平日	〇
曇	低	休日	〇
雨	高	平日	×
曇	低	平日	×
晴	低	平日	×
曇	低	平日	×

さらに分割

- 休日
  - 全て〇　⇒　停止
- 平日
- 属性：天気
  - 曇：3件．〇:1，×:2
    - Ent = 0.918
  - 晴：2件．〇:1，×:1
    - Ent = 1
  - 雨：1件．
    - Ent = 0
  - 加重平均：0.792
- 属性：気温
  - 高：3件．〇:2，×:1
    - Ent = 0.918
  - 低：3件．〇:0，×:3
    - Ent = 0
  - 加重平均：0.459

◎気温で分割

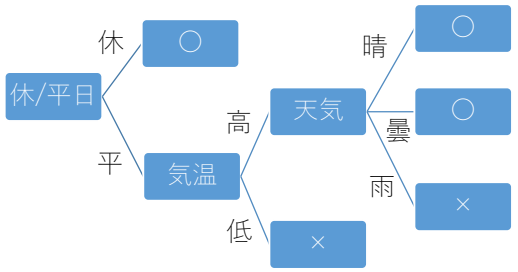
休日

天気	気温	売上
曇	高	〇
晴	低	〇
雨	低	〇
曇	低	〇

平日

天気	気温	売上
曇	高	〇
晴	高	〇
雨	高	×
曇	低	×
晴	低	×
曇	低	×

学習結果

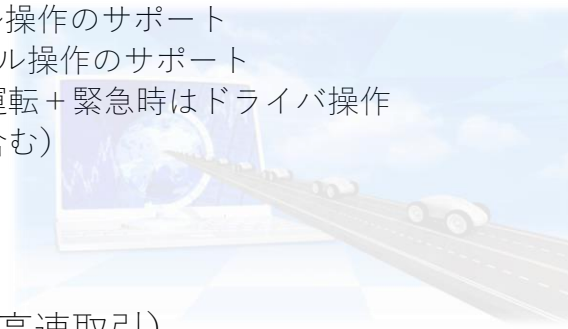


データマイニングとAIの関係

- 技術的には知識獲得手法
  - ビジネス応用の観点からの研究開発
- AI冬の時代に活発化
  - 「知識の時代」の課題：知識獲得のボトルネック
  - トイ問題：AIの扱う問題はいつも小規模・単純
  - 使える技術としての研究：第3次ブームにつながる成果

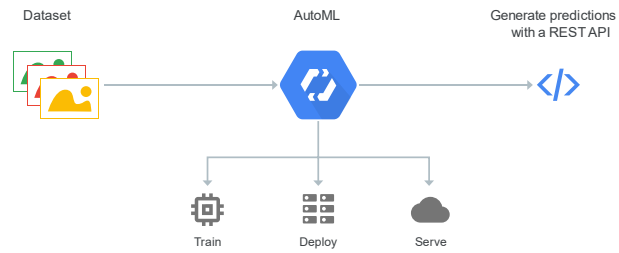
AI=自動化？

- 自動運転：5段階
  - ハンドル OR ブレーキ・アクセル操作のサポート
  - ハンドル AND ブレーキ・アクセル操作のサポート
  - 特定の場所（高速道路）で自動運転＋緊急時はドライバ操作
  - 特定の場所で自動運転（緊急時含む）
  - 完全自動運転
- 将棋・囲碁
  - 人間と対戦・勝利
- HFT（High-Frequency Trading: 高速取引）
  - ミリ秒単位での株の自動売買



AutoAI

- Google AutoML
- H2O.ai
- DataRobot
- Auto-sklearn
- Amazon SageMaker
- ワークフローの自動化
  - 準備：データ獲得・前処理・feature engineering
  - モデル構築：選択・ハイパーパラメータ調整，性能検証
  - デプロイ：モニタリング，改良含む



https://cloud.google.com/automl

AIとIA

- AI: 人間の知能の再現
- IA (Intelligence Amplifier): 人間の知能を増幅・支援
  - D. C. Engelbart：マウス，ハイパーテキスト
  - B. Shneiderman：情報可視化
- AI vs. IA? それとも AI ⊃ IA?

人工知能学会 山田誠二 会長が解説、「AIで人間の仕事が奪われる」は間違い

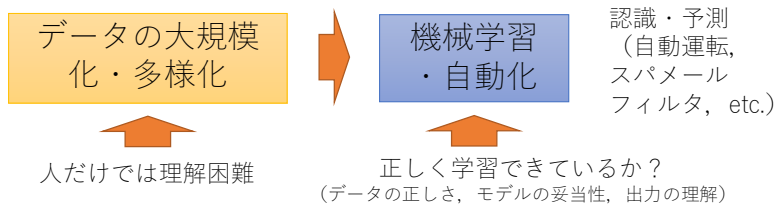
ディープラーニングが第3次AIブームを牽引し，さまざまなビジネス領域での活用が議論される．AIが真に普及するには「人間と人工知能の建設的な協調の議論が欠かせない」と語るのが，日本のAI研究の第一人者，人工知能学会会長の山田誠二氏だ．山田氏は，かつて産業革命期にイギリスの労働者が起こした機械排斥運動「ラッダイト運動」になぞらえ，「AIによって人間の仕事が奪われるのは誤った認識」だとして，両者の得意分野を相互に補うことが望ましいと提言した．

ビジネス+IT, 2017/5/22
https://www.sbbit.jp/article/cont1/33609

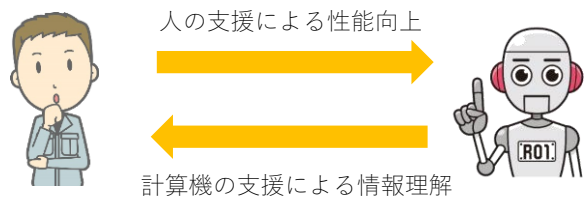
人と協調するAI

- 対話エージェント
  - ロボットによる観光案内，高齢者支援
  - チャットによる接客サービス
- 情報推薦
  - 計算機からのお薦め
  - オンラインショッピング
- 情報可視化
  - 人の優れた視覚の利用
  - 「見えにくいものをわかりやすく」

情報活用のキーワード



人と計算機のより良い協調関係の実現



データマイニングにおける人とAIの協調

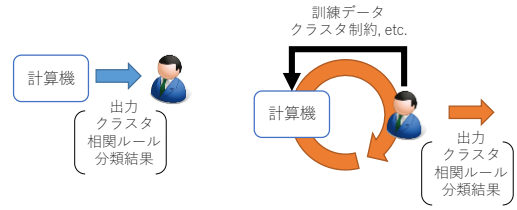
- AutoAIに対するデータサイエンティストの意見[Wang19]
  - 20名のデータサイエンティスト（IBM）に対するインタビュー調査
  - AutoAIをどのように活用できるか
- 意思決定者（クライアント）への説明可能性
  - クライアントが理解可能なモデル・信頼できる結果：仲介者としてのデータサイエンティストの役割
  - ベストのモデル ➡ 要求を満たすモデル
- AutoAIの獲得モデル ➡ 叩き台としての利用
  - 分析計画策定の支援
    - 行うべき分析，ラベル付けすべきデータ，吟味すべき特徴，評価すべきモデル
  - 教育用途：得られたモデルから学び，改良へ

データサイエンスにおけるAIとの協調

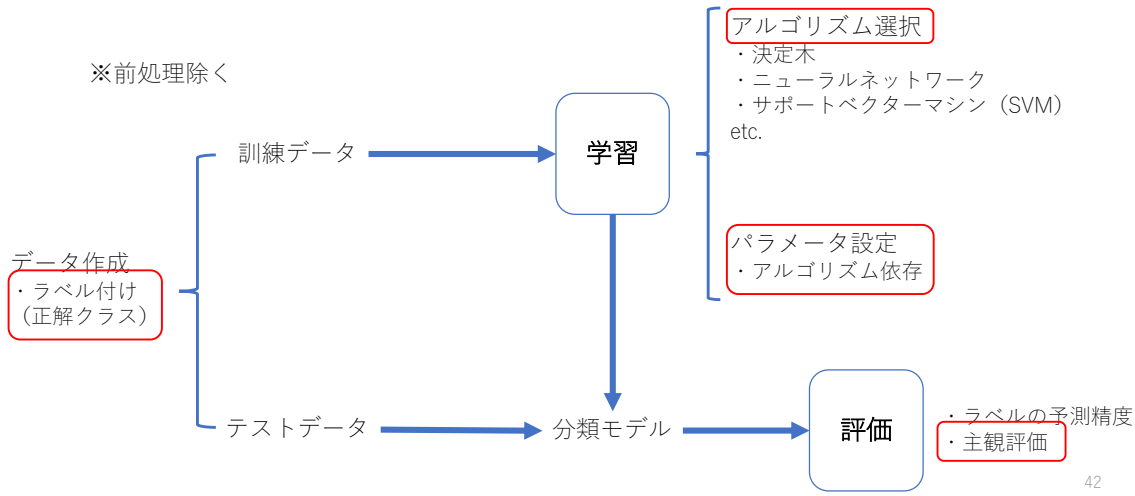
- Human in the loop
  - 半教師あり学習
  - インタラクティブなアルゴリズム
- 情報可視化
  - Visual Analytics

HITL: Human in the Loopとは

- 人間参加型：人間が計算機プロセスの中に入る
  - 人間と計算機の相互作用
  - 応用
    - シミュレーション　／　**データ分析**　／　**機械学習**
- 目的
  - 訓練（シミュレーション）
  - 人間の限界を克服（データ分析）
    - 作業記憶の限界，バイアス，疲労
  - 学習精度向上（機械学習）
    - 人間による判断・知識の導入
    - オーバフィッティングの軽減



教師あり学習のプロセス



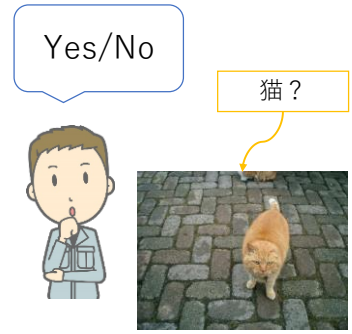
ラベル付け

- 機械学習（特に深層学習）は大量の訓練データが必要
- 世の中に大量に存在するデータ：ラベルなし
  - ラベル＝目的変数（所属クラス）： ground truth data
  - ラベル付与のコストが課題
- ラベル付けのアプローチ
  - 半自動ラベリング
    - クラウドソーシング (Crowdsourcing): 人数で解決
    - アクティブラーニング（active learning): ドメイン専門家の力
  - 自動ラベリング： ground truthなし
    - データプログラミング（Data Programing） [Ratner2016]：確率的ラベルの生成・それに基づく学習
    - Co-labeling [Li2012, Xu2016]：不確かなラベルからの学習

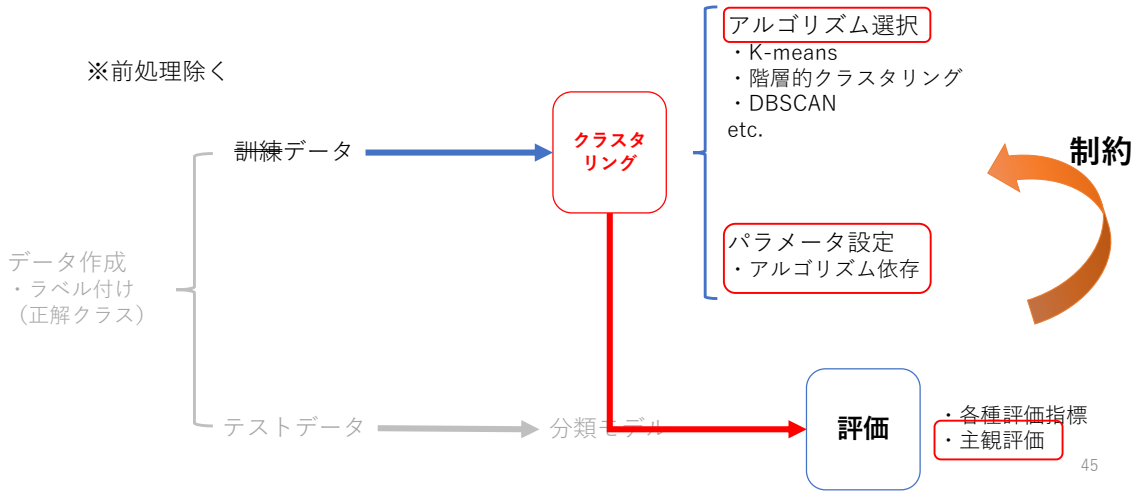
クラウドソーシングによるラベル付け

- ImageNet*
  - 代表的画像認識用データセット
  - WordNet （単語辞書）の階層と対応
  - 画像数：14,197,122
  - クラス数：21,841
- クラウドソーシングの利用
  - Amazon Mechanical Turk利用
  - ワーカ数：49,000
  - 複数ワーカが判断 ➡ 一定以上の確信度で採用
    - 確信度：物体，Yesの人数で決定

上位階層例	**動物, 装置, 食べ物, 植物, 人**
下位階層例	**ホタル, 論理回路, ベーグル, 気象学者**



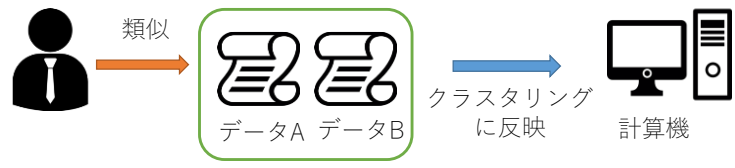
教師なし学習のプロセス



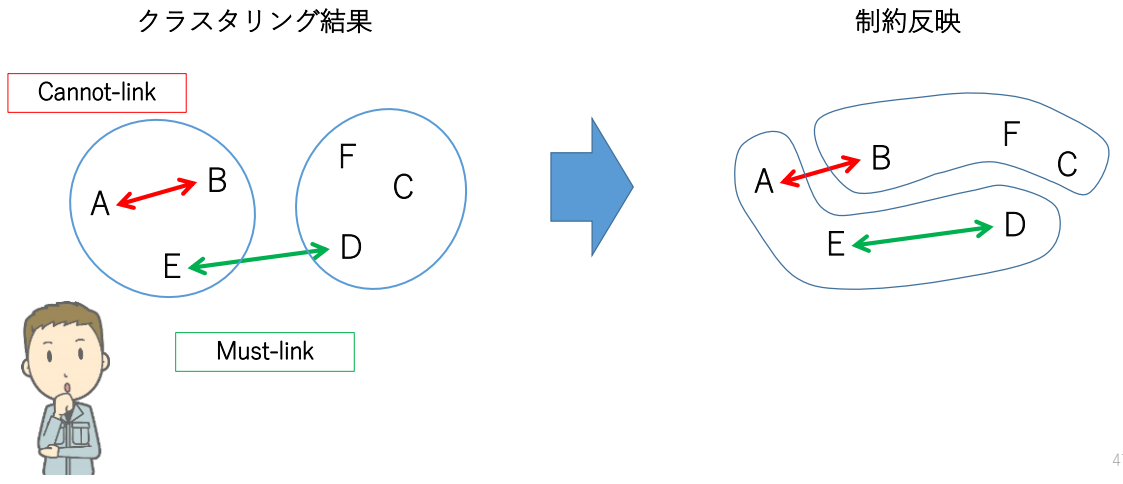
制約付きクラスタリング

- 制約：人間の判断を計算機に伝える役割
  - 一部データに制約　⇒　残りのデータも分類
- 制約の与え方
  - クラスタの大きさ・密度
  - ラベル：同じラベルのデータ＝同じクラスタに
  - 対制約：データのペア（対）に対する制約

Must-Link:
　同じクラスタに
Cannot-Link:
　別のクラスタに

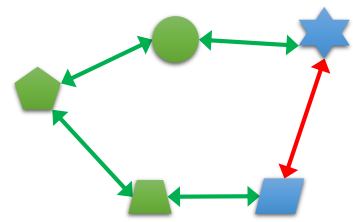


Must-link, Cannot-link



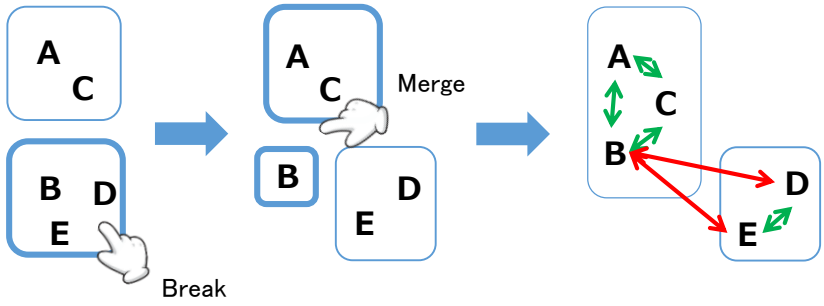
制約付与における課題

- ユーザ負荷の問題
- 制約矛盾の可能性
- 解決策：制約の一括付与
  - 1回の操作で複数の制約を生成



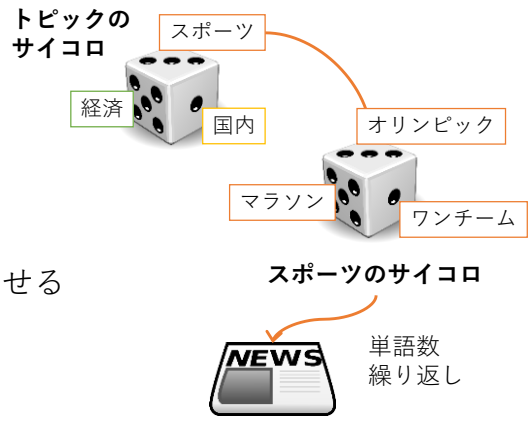
制約の一括生成

- グルーピング操作を解釈
  - 同一グループ内オブジェクトにMust-link付与
  - 別グループになったオブジェクト間にCannot-link付与



インタラクティブなトピックモデリング

- トピックモデリング
  - 教師なし学習の一種
  - レビュー（口コミ）の分析
    - 何について語られているか？
    - 魅力・課題の分析，価値観の分析
- Human in the loop:
  - 自動処理：人間の判断との違い
  - 一部の修正 ➡ 残りは計算機に任せる
  - トピックモデリングへの導入
    - 単語とトピックの関係修正など



オリンピック　マラソン　インバウンド

Human in the Loopの課題

- ドメイン専門家　≠　AIの専門家
  - 学習モデル・可視化手法に対する知識なし
  - 結果の解釈困難
  - パラメータ調整困難
- 対策
  - 専門家が理解しやすいモデルの利用
    - Prophet :専門家が理解しやすい時系列データのモデル[Taylor17]
  - セマンティックインタラクション [Endert12]
    - Visual Analytics

Visual Analyticsの課題と同じ

おわりに：人と計算機の協調に向けて

- （従来）向上した計算機能力の活用
  - より高速に，大規模に，小型に…
  - 人間が情報活用のボトルネックに
- 解決策：人と計算機の「役割分担」の見直し
  - 人工知能：知的処理の代行
  - 可視化：ボトルネックの解消
  - 協調：Human in the loop