機械学習とデータマイニングの基礎
- 概論 -

2023年10月-11月

鷲尾 隆
大阪大学産業科学研究所

本講義の構成概観

- 機械学習とデータマイニングの概論(鷲尾)
- より先端的な機械学習やデータマイニングの技術(原・ホーランド)

お話の概要

- データとは？
  - データの定義（データ・情報・知識）
  - いろいろなデータの形式
- 機械学習とデータマイニングとは？
  - 機械学習とデータマイニング研究の歴史的変遷
  - 機械学習とデータマイニングとは何をする技術か
  - 機械学習とデータマイニングと他技術との関係
  - 機械学習とデータマイニング技術の俯瞰
  - 世界の基礎研究コミュニティ
- 機械学習とデータマイニングツールの現状
- 代表的基礎技術と適用事例
  - 決定木技術
  - バスケット分析技術
- 機械学習とデータマイニング技術の高度化例
  構造データマイニング

データとは？

各種辞典は何と説明しているか？

●三省堂大辞林

① 判断や立論のもとになる資料・情報・事実。

② コンピューターの処理の対象となる事実。状態・条件などを表す数値・文字・記号。

●goo国語辞書

① 物事の推論の基礎となる事実。また、参考となる資料・情報。

② コンピューターで、プログラムを使った処理の対象となる記号化・数字化された資料。

データとは？

- IT用語辞典

① データとは、何かを文字や符号、数値などのまとまりとして表現したもの。人間にとって意味のあるものや、データを人間が解釈した結果のことを情報と呼ぶ。

② 単にデータといった場合、ITの分野では特にコンピュータが記録・処理できる形式になっているもののことを指す。また、コンピュータが保存しているもののうちプログラム以外のもの、プログラムによって処理されるもの、という意味でデータという場合もある。

- Wikipedia

① データとは、基礎的な事実や資料をさす言葉。情報処理や考察によって付加価値を与える前提で集められており、基本的に複数個の事象や数値の集合となっている。

② 電子データは、コンピュータ内にあるか、コンピュータに取り込める形になったデータである。

データとは？

資料・情報・事実
何かの文字や符号、数値などのまとまり(集合)による表現
(伝達、解釈、処理などに適するよう形式化、符号化されたもの)

人間による解釈・推論・判断・立論，情報処理や考察による利用を前提に集められたもの

両方に当てはまるもの

電子データとは？

コンピューター処理の対象であるデータ
コンピュータが記録・処理できる形式になっているデータ
コンピュータ内にあるか、コンピュータに取り込める形になったデータ

参考：情報とは？

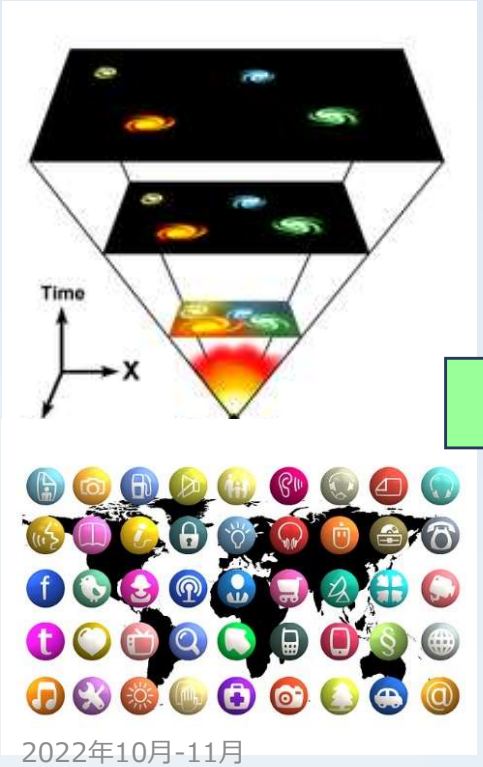
- 三省堂大辞林
  事物・出来事などの内容・様子。また、その知らせ。
- goo国語辞書
  ある物事の内容や事情についての知らせ。
- IT用語辞典
  - 物事の事情を人に伝えるもの。人が知覚したときに何らかの意味を想起させ、思考や行動に影響を与えるものを指し、人にとって意味を成さない雑音やランダムなパターンをも含む「データ」とは区別される。
  - ただし、情報科学・情報理論の分野では、情報の意味や価値判断の側面をひとまず捨象して、量的側面からその伝達や保存、変換について検討しており、この場合の「情報」は基本的にはデータと区別されない。
- Wikipedia
  - あるものごとの内容や事情についての知らせのこと。
  - 文字・数字などの記号やシンボルの媒体によって伝達され、受け手において、状況に対する知識をもたらしたり、適切な判断を助けたりするもののこと。
  - 生体が働くために用いられている指令や信号のこと。

参考：知識とは？

- 三省堂大辞林
  - ある物事について知っていることがら。
  - ある事について理解すること。認識すること。
- goo国語辞書
  - 知ること。認識・理解すること。また、ある事柄などについて、知っている内容。
  - 考える働き。知恵。
- Wikipedia
  - 認識とほぼ同義の語である。知識は主に認識によって得られた「成果」を意味するが、認識は成果のみならず、対象を把握するに至る「作用」を含む概念である。
- オックスフォード英語辞典
  - 経験または教育を通して人が獲得した専門的技能。ある主題についての理論的または実用的な理解。
  - 特定分野または一般に知られていること。事実と情報。
  - 事実または状況を経験することで得られた認識または知悉。

情報･データ･知識の関係

物事・出来事・事情



表現に変換

情報
内容・様子に関する知らせ

人間

データ
人間の情報処理利用を前提に集めた文字・符号・数値などの集合による表現

認識・解釈

知識
人間の認識により知っている事柄

いろいろなデータの形式

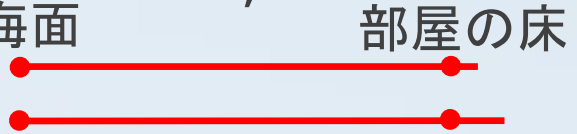
- データは形式を持つ変数値，変数値の関係，変数の関係からなる集合
  - 変数値の形式
    - 数値
      - 整数　　-10,-4,0,1,15,600
      - 実数　　-100.5, -1/7, 0.2, 1/3, 113/11
      - 複素数　　-5.2+2.3i, 4.5-10i
      - 正数･負数　　-11.4, -8.0, -0.5　　0.1, 1.5, 17.6
    - 記号（ラベル）
      - 山田太郎，佐藤花子
      - 良，悪
      - 大，中，小
      - 男性，女性

いろいろなデータの形式

- データは形式を持つ変数値，変数値の関係，変数の関係からなる集合
  - 変数値の関係
    - 比に意味がある比例尺度（絶対原点を持つ）
      価格，質量など

２００円は１００円の２倍の価値を持つ

    - 差(距離)のみ意味がある間隔尺度（絶対原点を持たない）
      摂氏温度，海抜高度など



100m, 1m　床からは2倍の高さ
101m, 2m　海面からは1.01倍の高さ

いずれも差は1m

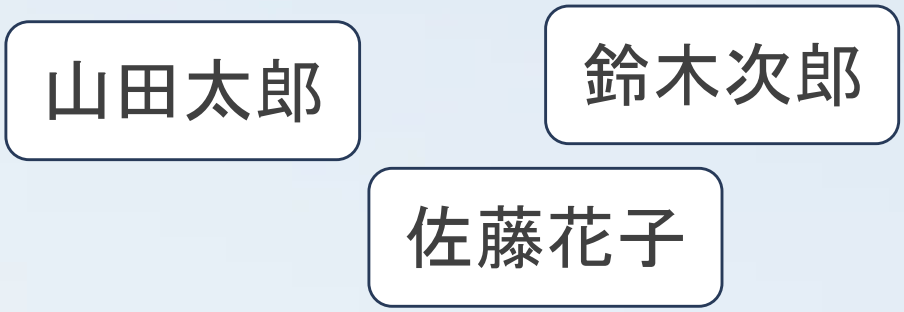
    - 順序のみ意味がある全順序尺度
      成績順位，光の色（赤～紫）など

高　周波数　低

いろいろなデータの形式

- データは形式を持つ変数値，変数値の関係，変数の関係からなる集合
  - 変数値の関係（つづき）
    - 部分的順序のみ決まっている半順序尺度
      乗り物の速さ

    - 区別することのみ意味がある名義尺度
      学籍番号，性別，氏名など



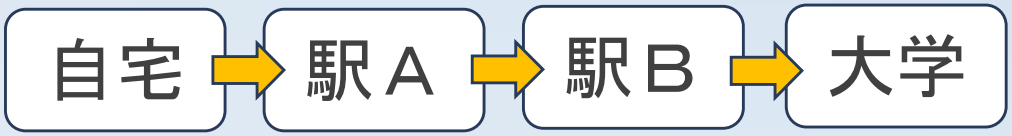
別人であることを区別できるが、それらの関係を表さない

いろいろなデータの形式

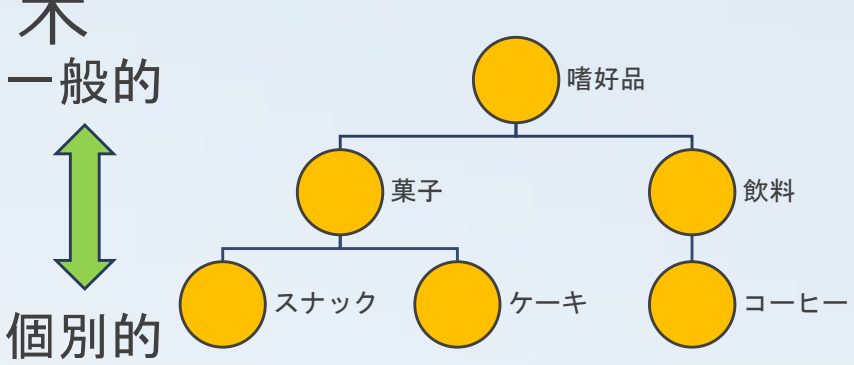
- データは形式を持つ変数値，変数値の関係，変数の関係からなる集合
  - 変数の関係
    - 定量的または定性的規則性

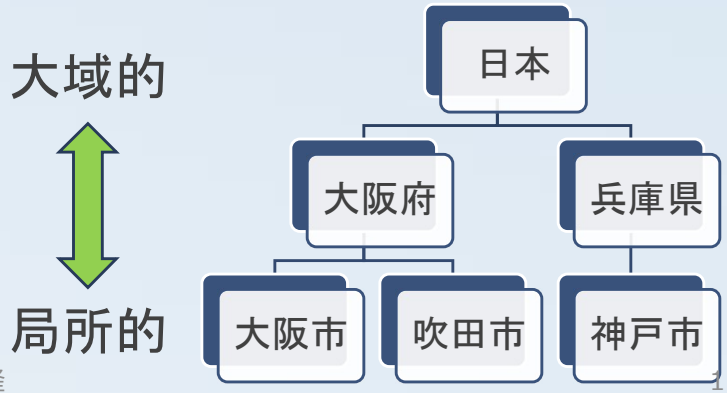
      コーヒーの値段=360円，コーヒーの値段+ドーナツの値段=540円

      気温，湿度について蒸し暑さは単調増加
    - 系列



    - 木

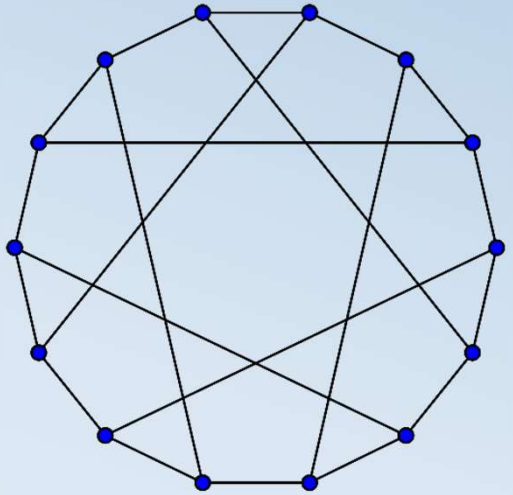


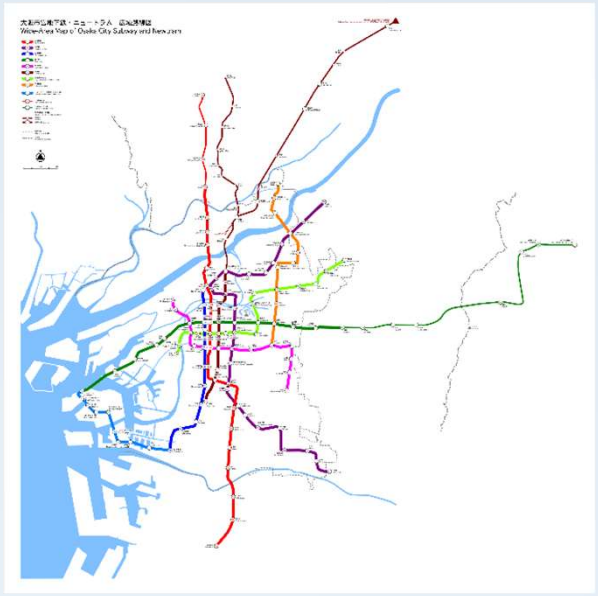


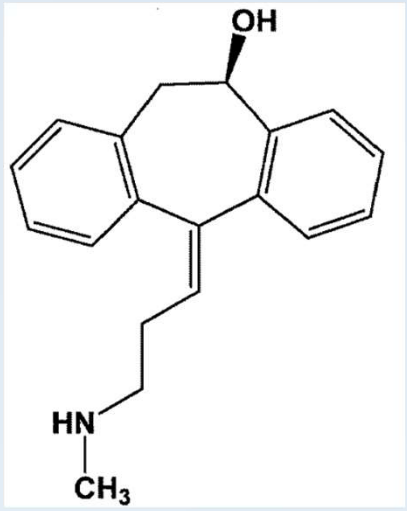
いろいろなデータの形式

- データは形式を持つ変数値，変数値の関係，変数の関係からなる集合
  - 変数の関係（つづき）
    - グラフ（ネットワーク）

$$\begin{array}{ccccccc} A & \xrightarrow{f} & B & \xrightarrow{g} & C & \longrightarrow & 0 \\ \downarrow{\scriptstyle a} & & \downarrow{\scriptstyle b} & & \downarrow{\scriptstyle c} & & \\ 0 & \longrightarrow & A' & \xrightarrow{f'} & B' & \xrightarrow{g'} & C' \end{array}$$





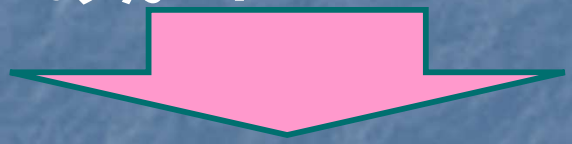


お話の概要

- データとは？
  - データの定義（データ・情報・知識）
  - いろいろなデータの形式
- 機械学習とデータマイニングとは？
  - 機械学習とデータマイニング研究の歴史的変遷
  - 機械学習とデータマイニングとは何をする技術か
  - 機械学習とデータマイニングと他技術との関係
  - 機械学習とデータマイニング技術の俯瞰
  - 世界の基礎研究コミュニティ
- 機械学習とデータマイニングツールの現状
- 代表的基礎技術と適用事例
  - 決定木技術
  - バスケット分析技術
- 機械学習とデータマイニング技術の高度化例
  構造データマイニング

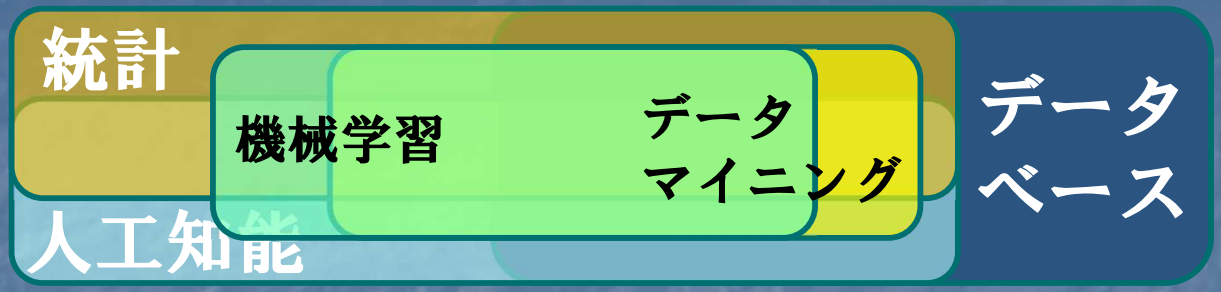
機械学習とデータマイニングとは？

- 人によって意味していることが違う。
  - 25年くらい前から知られるようになり、書店にいくとExcelでデータマイニングとか、いろんな本が並んでいる。でも中には相関解析、重回帰分析、判別分析など、従来の統計手法による分析方法を解説しているだけのものもある。。。。
  - データベースや検索エンジンで情報検索して、いろんな情報を見つけたり、傾向を調べることも行われている。。。。
- 機械学習やデータマイニングは、統計分析や情報検索と変わらないのか？



- まずは，この素朴な疑問についてお話したい。
  - ただし、人によって意見が違う部分もあるので、一部私見が入る。
  - 日進月歩で技術が発展し、その概念自体が急速に変化している.

機械学習とデータマイニングの歴史



統計やデータベースを基礎。
人工知能技術の中核。
両者はかなり共通する。

長い長い学問の歴史の産物。最近急に出てきたのではない！

統計の歴史

17世紀 ウィリアム・ペティの『政治算術』
18世紀 ジュースミルヒの『神の秩序』
19世紀 カール・フリードリヒ・ガウス、カール・ピアソン
20世紀 イェジ・ネイマン、エゴン・ピアソン

機械学習とデータマイニングの歴史

1956年　　ダートマス会議＝人工知能(機械学習)の創始
1960-70年代 パーセプトロン・重回帰・パターン認識
1980年代　決定木・ニューラルネット＝機械学習の創始
1990年代　データマイニングの創始
2000年代　統計的機械学習の創始
2010年代　ビッグデータ解析の創始・ディープラーニング

高校で習う数学・統計をはじめとして、高等数学の固まり。

地道な学問と技術

機械学習とデータマイニングとは？（その2）

- 現実の不均一な（ムラのある）膨大なデータ
  - 企業やネットワークのデータなどは、様々な条件や対象のデータ。
    - 世論調査：都市と地方、関東と関西では意見も違う。。。
    - 商品売上：顧客年齢や性別、地域で購入商品や購入金額が違う。。。
  - 多くの統計手法のデータ分布均一性の仮定が成立しない。
    - 標本抽出して解析、全データを1つの統計分布で分析など。。。
    - 一様に砂鉄が含まれる砂浜。

- 実際に必要なデータ傾向分析
  - 異なる傾向のデータに分け、それぞれ傾向分析。
  - 場所によって鉄鉱脈があったりなかったり。あたかも発掘現場。
- 1996年 U. Fayadd: Data mining
  - 膨大で不均一なデータから有用な知識を掘り出す方法の研究
    - データをある程度均質な分布のセグメントに分解
    - 各セグメントについて適切なモデリングや解析

機械学習とデータマイニングとは？（その3）

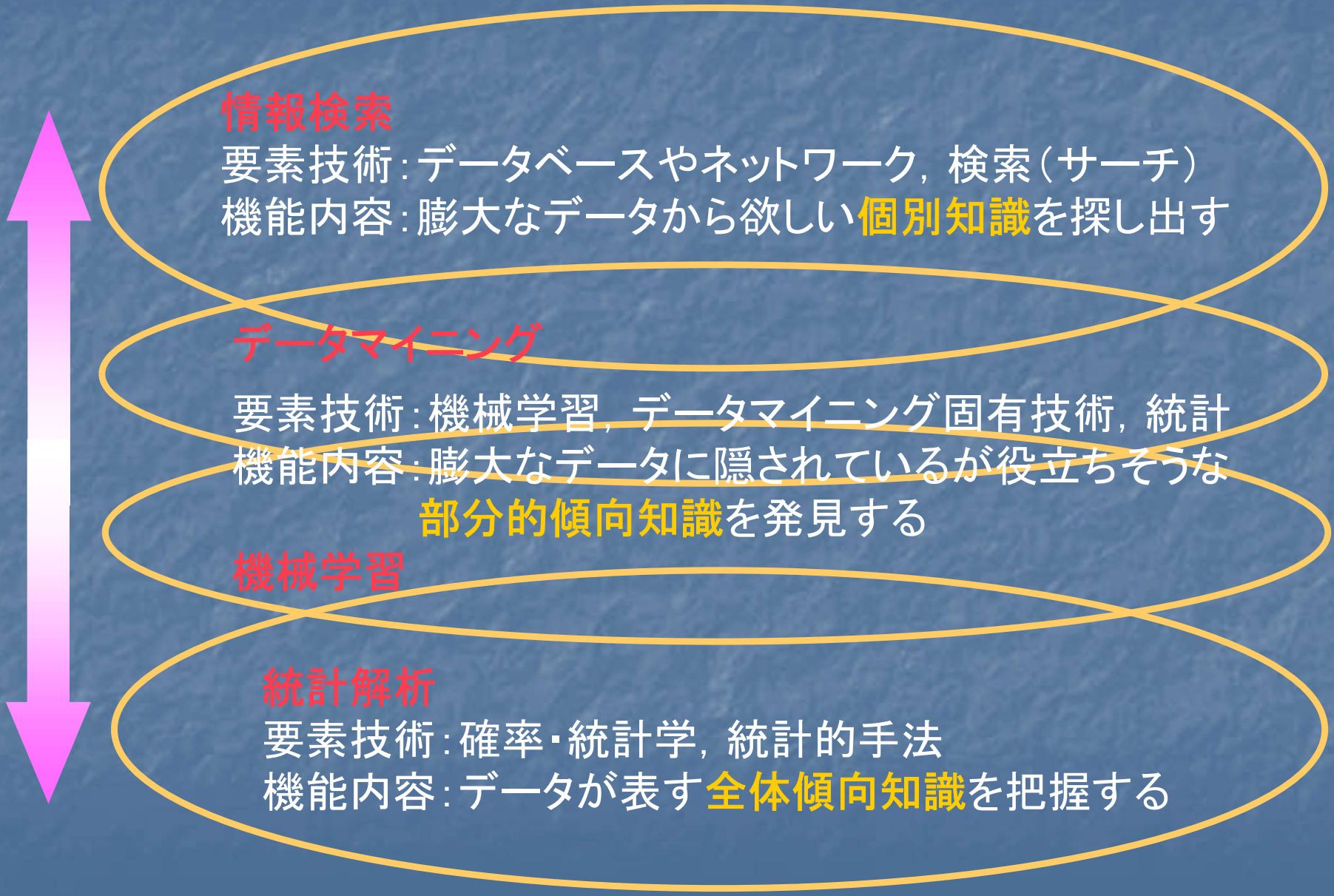
- ただし、分解されたセグメントから分析によって分かる傾向は、見掛け上のものかも知れない。
  - 統計誤差：セグメントデータが少な過ぎ。
  - 系統誤差（バイアス）：セグメントデータが特定の性質に偏り過ぎ。

データマイニング

•データから最終結論を知識として取り出すとは限らない。
•多くの場合、**可能性のある仮説を知識**として発掘する。

- データマイニングは、多くの場合、他の仮説検証方法と組み合わせて用いられる。
  - 対象セグメントに関してもっとデータを集めて統計精度を高める。
  - 実験や実地調査を行って、本当かどうかを直接確かめる。

データマイニングと情報検索，統計との関係



機械学習とデータマイニングの多様性

- 知識発見：
  - あくまでも導出結果から，人間が解釈して知識として重要なものを選び取ることが目的．

    通常，統計的有意性検定を課さないで，粗いフィルタリングに留める．
    - 統計的有意性と人間の解釈可能性が必ずしも一致するとは限らない．
    - 人間がデータを追加収集し検討・解析したり，他の実験や専門的知見によって検証する余地を残す．
    - 人間が必要に応じて統計的有意性検定を実施．

機械学習とデータマイニング
データマイニングの多様性

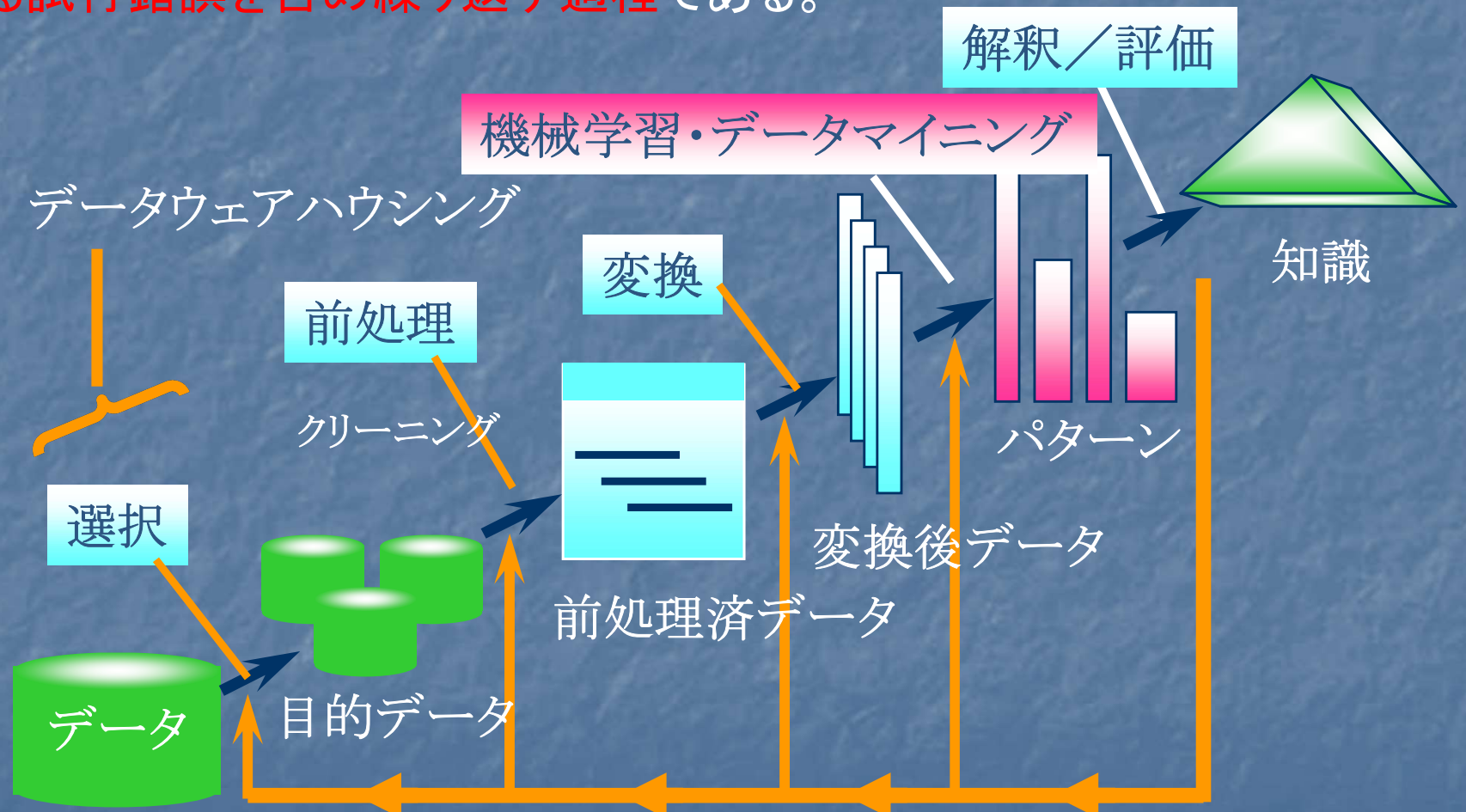
- 最適予測:
  - 統計的手法による予測と同様，予測結果について，各種精度指標による性能評価やクロスヴァリデーションによる予測の妥当性，安定性検証などが行われる.
  - 予測手法の性質上，過学習（オーバーフィッティング）を生じない場合には，クロスヴァリデーションを省略する場合もある.

機械学習とデータマイニングの多様性

- 対象データ:
  - データ数:
    - 数千から数千万個までという広い範囲にまたがる.
  - データの形式:
    - 数値, ノミナルな記号, 天体写真のような画像, ホームページのような自然文やタグ・リンク付きハイパーテキスト, 遺伝子配列のような系列, 半順序を表す木構造, 化学式のようなグラフなど
  - データ品質:
    - ノイズの多寡, 欠測値がの多寡, 実験や調査の制約上の不可避的に強いデータバイアス, 解析目的が遂行可能な品質や内容のデータかどうかは事前に分からないことが多い.
  - データ前処理:
    - 型通りの手法に適用できるように多大な労力を要する. 形式変換, データ項目の取捨選択, 欠測値の補完(人間が介在する場合もある), 他データからの新たなデータ項目生成など.

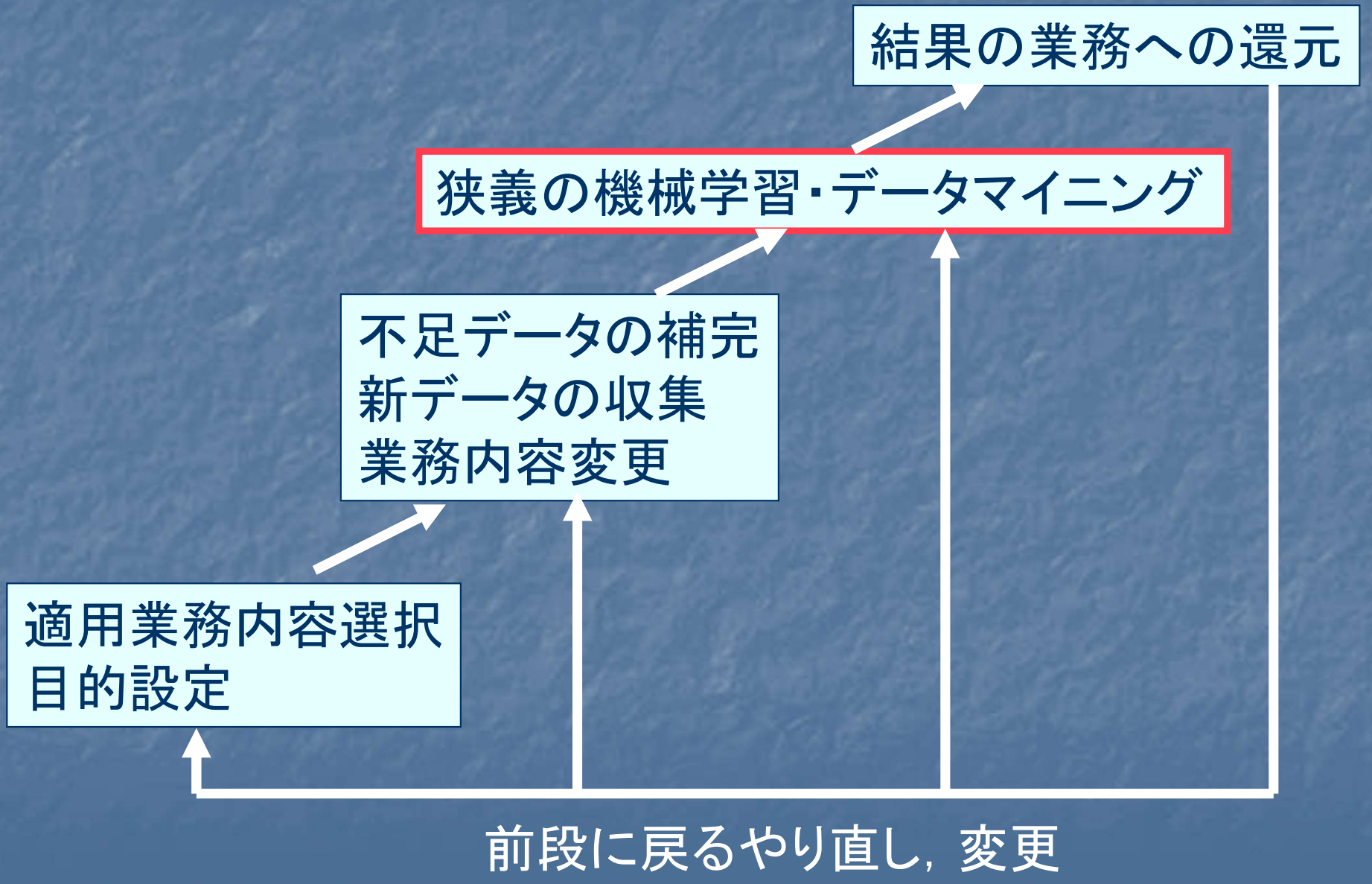
機械学習とデータマイニングのとは？（その4）

- 多くのデータマイニング技術は**データ解析者を支援のためのもの**であり、解析を自動化するためのものではない。
- 目的に応じた知識が得られるまで、様々な**データ処理や解析を人手による試行錯誤を含め繰り返す過程**である。



データマイニングの解析過程 [U.Fayyad(1996)]

更に機械学習･データマイニング（知識発見）とは何か？ ＝総合工学の1つ
業務改革や組織論までもが関係する



機械学習とデータマイニングとは？（その5）

- 機械学習とデータマイニングは一種の情報総合工学
  - 数千種類を超える前処理、変換、解析、解釈・評価手法の総称
  - 機械学習（人工知能）、統計数理、データベース、情報検索など、種々の情報処理技術から成り立つ。
- 技術内容を益々膨らませている。
  - テキストマイニング
  - Webマイニング
  - バイオインフォマティクス
  - パターン認識
  - 信号処理関連など

機械学習とデータマイニング技術の俯瞰（その1）

データ解析技術（殆どが完全自動解析ではなく、人間の介在が必要。）

- **分類技術**：データ中の各事例を分類する規則や数式を見出す。

  決定木系技術、深層学習・ニューラルネット系技術、ベイジアンネット系技術、統計的判別分析系技術、 サポートベクターマシン系技術、エマージングパターンルール系技術、アンサンブル学習系技術など
- **クラスタリング技術**：データを似た事例のグループに仕分けする。

  k-means系技術、密度ベース系技術、樹状図系技術、ニューラルネット系技術、カーネル関数系など
- **バスケット分析技術**：事例に共起する条件を見出す。

  Apriori系技術、Pattern Growth系、LCM系技術、その他多数
- **構造パターン解析技術**：系列や木構造、グラフ構造データから頻出パターンを見出す。

  系列マイニング系、木構造マイニング系、グラフマイニング系、カーネル関数系など
- **統計的関数分析技術**：データ項目間の関係やその中で主要な関係を見出す。

  重回帰分析系、ロジスティック回帰系主成分・因子分析系、独立成分分析系など

各々の技術について、代表的なものだけで数百種類ある。

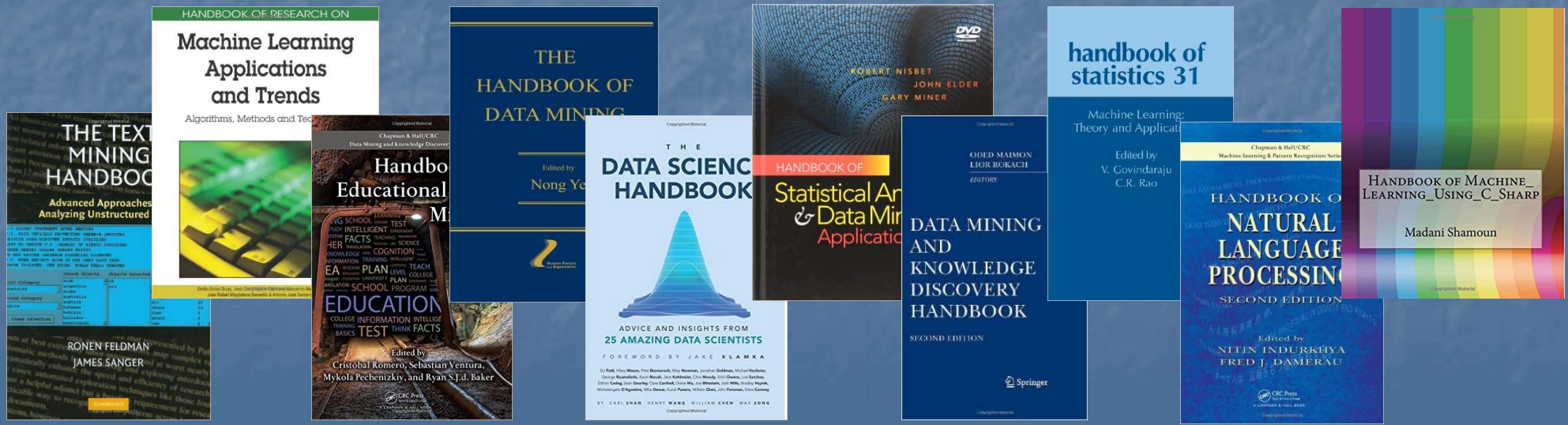
機械学習とデータマイニング技術の俯瞰（その2）

- **データ前処理技術**（殆どが完全自動解析ではなく、人間の介在が必要。）
  - 属性選択技術
    - データから解析目的に必要性の高い項目属性を選択する。
  - 属性生成技術
    - データから解析目的に必要性の高い項目属性を合成する。
  - 事例選択技術
    - データから解析目的に必要性の高い事例を選択する。
  - 事例生成技術
    - データから解析目的に必要性の高い事例を合成する。
  - 数値データ離散化技術
    - 数値データを解析目的に必要な記号データに離散化する。
  - その他（細かな処理技術は多数）
- **データ可視化技術**：複雑な発掘結果を人間に分かりやすく表示。

上記各技術について、代表的なものだけで数十～数百種類ある。

データマイニング・機械学習技術は膨大

- 技術があまりに多岐なため、百科事典ともいうべき**ハンドブックが多数刊行**されている。
  - Handbook of Data Mining and Knowledge Discovery
  - Data Mining and Knowledge Discovery Handbook
  - Handbook of Statistics, Volume 24: Data Mining and Data Visualization
  - Handbook of Data Mining
  - Handbook of Educational Data Mining
  - The Text Mining Handbook
  - Handbook of Research on Machine Learning Applications and Trends
  - Handbook Statistical foundations of machine learning



世界的基礎研究の中心コミュニティ(その1)

- 機械学習とデータマイニング基礎研究の中心は主要国際会議
- データマイニング中心
  - SIG-KDD:米国ACM主催，データマイニング分野で頂点の国際会議
  - ICDM:米国IEEE主催，SIG-KDDに並びデータマイニング分野の頂点
  - SDM:米国SIAM主催，上記2つほどではないが，発表論文は広く読まれる。
  - (PAKDD:アジア系，比較的高レベルの論文が集まる。)
- 機械学習中心
  - ICML:米国系，機械学習分野で頂点の国際会議
  - PKDD/ECML:欧州系，上記ほどではないが非常に高レベルの論文が集まる。
  - (ACML:アジア系，上記ほどではないが比較的高レベルの論文が集まる。)
- 統計・数理モデル中心
  - NeurIPS:米国系，ニューラルネットや数理モデル分野で頂点の国際会議
  - AISTATS :米国・欧州系，確率統計ベースの非常に高レベルの論文が集まる。
  - UAI:米国系，確率統計ベースの人工知能分野で頂点の国際会議

世界的基礎研究の中心コミュニティ(その2)

- データベース中心
  - SIG-MOD：米国ACM主催，データベース分野で頂点の国際会議
  - VLDB：米国系，SIG-KDDに並びデータベース分野で頂点の国際会議
- 情報検索中心
  - SIG-IR：米国ACM主催，情報検索分野で頂点の国際会議
  - CIKM：米国ACM主催，情報検索やデータマイニング分野で非常に高レベルの論文が集まる。
- 人工知能中心
  - IJCAI：米国系だが世界持ち回り開催，人工知能分野で頂点の国際会議
  - AAAI：米国AAAI主催，IJCAIに並んで人工知能分野で頂点の国際会議

- その他、テキストマイニング、Webマイニング、バイオインフォマティクス、パターン認識など、各分野の頂点国際会議でも高レベルのデータマイニング論文が発表されている。

お話の概要

- データとは？
  - データの定義（データ・情報・知識）
  - いろいろなデータの形式
- 機械学習とデータマイニングとは？
  - 機械学習とデータマイニング研究の歴史的変遷
  - 機械学習とデータマイニングとは何をする技術か
  - 機械学習とデータマイニングと他技術との関係
  - 機械学習とデータマイニング技術の俯瞰
  - 世界の基礎研究コミュニティ
- 機械学習とデータマイニングツールの現状
- 代表的基礎技術と適用事例
  - 決定木技術
  - バスケット分析技術
- 機械学習とデータマイニング技術の高度化例
  構造データマイニング

機械学習・データマイニングツールの現状（1）

現状の市販ツールベンダー

IBM, 数理システム, スポットファイア, マスワークスなど

現状の市販ツールのサポート手法

深層学習：CNN, RNN, Auto Encoder, Deep Belief, Deep Boltzman

ニューロ：BP, MLP, RBF

決定木： ID3, C4.5, C5.0, CART, CHAID, QUEST,
Pseudo Decision Tree, Option Tree, ファジイ決定木,
２次元領域の抽出決定木，機能拡張決定木

クラスタリング：コホーネン，K-means，Ward法，
コンドルセ手法，概念クラスタリング

相関ルール：Apriori, Generalized Rule Induction,
順序アソシエーション，複数時系列

統計的手法：重回帰分析，ロジスティック回帰分析，
判別分析，主成分・因子分析, ベイジアンネット

テキストマイニング，Concept Base Search，記憶ベース推論

機械学習・データマイニングツールの現状（2）

オープンソースフリーウエア（OSでいうLinux）

- Python　汎用のプログラミング言語であるが、少ない行数で書ける特徴を生かし、データマイング、機械学習、統計解析の各種アルゴリズムを実装したコマンドライブラリが整備されている。
- TensorFlow, PyTorch, Microsoft Cognitive Toolkit: Google, Meta, Michrosoftなど、民間企業が深層学習の普及を狙って無料公開している深層学習用解析ツール
- Weka　ニュージーランドワイカト大学のチームが開発。業務用ではなく、データマイニング研究者やデータマイニング試用時の小規模検証向き。
- R　元商用のS-Plusというデータマイニングソフトをベースにフリーウエア版にしたもの。中規模データまで扱える。統計解析中心。

オープンソースフリーウエアのサポート手法

若干インターフェースが落ちるが機能は商用版とほぼ同じ

- 深層学習：CNN, RNN, Auto Encoder, Deep Belief, Deep Boltzman
- ニューロ：BP, MLP, RBF
- 決定木：ID3, C4.5, C5.0, CART
- クラスタリング：コホーネン，K-means法，概念クラスタリング
- 相関ルール：Apriori, Generalized Rule Induction, 複数時系列, グラフマイニング
- 統計的手法：重回帰,ロジスティック回帰,判別分析,主成分・因子分析,ベイジアン
- テキストマイニング，Concept Base Search，記憶ベース推論

用語による混乱の問題

- 各学問領域固有の用語の問題
  機械学習・データマイニングにも固有の用語が多数.
  - 属性　　:特徴量や説明変数
  - クラス　:目的変数
  - 支持度　:出現頻度(確率と頻度の違いは微妙), ...
- 外来語の和訳の問題
  - AssociationとCorrelation:相関
    - Correlation:物事の間に存在する関係
      物理的関係:Aが起こればBが起こる, Aが起きなければBも起きない.
    - Association:記憶や想像の中で何らかの繋がりをもつこと
      共起の観察関係:
      AとBは共に起き易い. Aが起きなくてもBが起きる可能性はある.
- 結果の取り扱いの問題
  - 一般にデータマイニングで得られる規則や傾向はあくまで仮説.
    客観的事実かどうか, 現実に当たって検証(統計的検証も含む)
    する必要がある.