In Silico Data
Miracles by the KY-methods

The Chem-Bio Informatics Society 情報計算化学生物学会

CBI学会2019年大会、2019年10月21日（月）、13：00－17：00

〈チュートリアル〉

計算毒性学と化学データサイエンスの基本

株式会社　インシリコデータ
湯田　浩太郎

本日のプログラム：

1．13:00-13:05:（5分） 挨拶：株式会社　インシリコデータ　湯田浩太郎

2．13:05-13:20（15分） ◆導入　計算毒性学と「化学データサイエンス」
計算毒性学でのコンピューター導入原理、二大毒性評価関連技術（化学多変量解析／パターン認識アプローチ、人工知能アプローチ）、データサイエンスから「化学データサイエンス」へ

3．13:20-13:50（30分） ◇第一部　計算機化学（Computer Chemistry）関連
化合物保存形式、化合物命名法、化合物検索（完全一致、部分構造、2・3次元構造検索、他）手法、一元一項対応串刺し検索、化合物の扱い（縮合多環、互変異性、立体／幾何異性）、化合物表記（ケトエノール、ニトロニトロソ、他）

4．13:50-15:20（90分） ◇第二部　化学多変量解析／パターン認識（ケモメトリックス（Chemometrics））関連
化学パラメーター、2／3次元パラメーター、種々データ解析手法、過剰適合、偶然相関、線形／非線形性、特徴抽出、最少サンプル数、最少パラメーター数、クラスポピュレーション、次元変換／圧縮／縮小、分類率／予測率、要因解析、オートスケーリング、アウトライヤー／インライヤー、解析信頼性指標（サンプル数／パラメーター数）、KY（K-step Yard sampling）法、パーセプトロン、バックプロパゲーション、遺伝的アルゴリズム、ファジー理論、内挿／外挿問題、他

＜15:20-15:40　休憩 20分＞

5．15:40-16:20（40分） ◇第三部　人工知能（Artificial Intelligence）関連
人工知能の歴史、ルールベース型人工知能、ニューラルネットワーク型人工知能、深層学習、サンプル数問題、要因説明問題、ルールのコンピューターへの組み込み、ネットワーク構造、LISP、FORTRAN、PYTHON、

6．16:50-17:00（30分） ◇第四部　計算機科学（Computer Science）関連
データベース理論、プログラミング言語、クラスター、クラウド、スーパーコンピューター、ネットワーク、WEB、他

7．16:50-17:00（10分） ◇討論および名刺交換会

◆化学多変量解析／パターン認識（ケモメトリックス（Chemometrics））関連キーワード

化学パラメーター
２／３次元パラメーター
種々データ解析手法
過剰適合
偶然相関
線形／非線形性
特徴抽出、
最少サンプル数
最少パラメーター数
クラスポピュレーション
次元変換／圧縮／縮小
分類率／予測率
要因解析
オートスケーリング
アウトライヤー／インライヤー
解析信頼性指標（サンプル数／パラメーター数）、
KY（K-step Yard sampling）法
パーセプトロン
バックプロパゲーション
遺伝的アルゴリズム
ファジー理論
内挿／外挿問題
他

□データ解析の現状

イントロのイントロ（１）

抽選機（データ解析ソフトウエア）回したら玉（解）が出た

その玉（解）は、**当たり？　はずれ？**

イントロのイントロ（２）

- データ解析ソフトに**データを入れれば必ず答えが出る**
  ⇒ これが**最も深刻な問題**です

勘違い：
⇒ 解が出たから**プロフェッショナル**だー
⇒ 一件落着で仕事完了。**仕事したなーー・・・、疲れたけど達成感**
⇒ 解を解析して、**次の仕事に向けた情報収集**しようか。
＊そんなことして大丈夫かなーー？？ **解の信頼性**は？？
⇒ データ解析手法を**数式レベルで展開**できるので間違うはずはない
＊データ解析手法の**基本原理**と**運用**は**別問題**です

□イントロのイントロ（３）

初心者：データ解析ソフトが動いて解が出てくるレベル

単にマニュアルに従って操作したら解が出た・・

＊素晴らしいですね、解が出てきたのですから。達成感あるでしょう・・・。

質問①単にラッキーだっただけではありませんか？

・数値データにミッシングデータが混ざっていなかった

・数値データの桁数は？

多少桁数が変わっていてもプログラムは解を出します

・0データの扱いはどのようにしましたか？

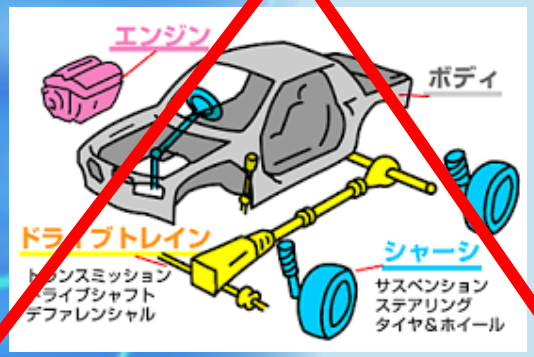
・分類／予測率、相関／絶対係数しか気にしてませんか？

○ベテラン（データサイエンティスト）：正しい答えを出す保証の有無

○個々の分野の研究者：出てきた解を正しく評価／討論できる力

日本チュートリアルの目指すところ

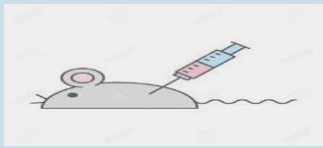
データ解析手法　基本原理等



データ解析関連技術

道路交通法、マップ、天候
道路（坂、砂、高速、砂利等）

適用分野



口本チュートリアルの目指すところ

データ解析手法 基本原理等

~~ニクラス分類~~
~~多クラス分類~~
~~重回帰~~
~~ニューラルネットワーク~~
~~バイナリーツリー~~
~~マッピング~~
~~クラスタリング~~
~~チャート分析~~
~~SVM~~
~~PCA~~
~~AdaBoost~~
~~ランダムフォレスト~~
~~その他~~

基本的な部分は説明

データ解析関連技術

○データ解析の適用限界
・個々の手法に依存
○サンプリング
・最小サンプル数
・クラスポピュレーション
・サンプリングプロトコル
・アウトライヤー、インライヤー
○線形／非線形問題
・空間合致と空間の再構築
・外挿と内挿
○パラメーター
・種類
・パラメーター選択
・0値の扱い
・スケーリング
○解析信頼性
・サンプル数／パラメーター数
・要因解析
・クロスバリデーション

適用分野

●創薬関連
・Q／T／ADME／P／SAR
・インシリコスクリーニング
・ドラグリストラクチャリング
・要因解析
●化合物毒性評価
・TSAR（構造毒性相関）
・毒性予測
・脱毒性
・化合物＆環境規制
●機能性化合物デザイン
・PSAR（構造物性相関）
●機器スペクトル
・スペクトル解析
・メタボロミクス
●バイオ解析関連
●医療との連携（画像等）
●その他

◇「ケモメトリックス」とは

ケモメトリックスとは：
計量化学（けいりょうかがく、chemometrics）とは、数理科学、統計学、機械学習、パターン認識、データマイニングなどの手法により、（広義の）化学分野における諸問題を解決しようとする分野である。

ウィキペディアより：https://ja.wikipedia.org/wiki/計量化学

＊ちなみに、chemometricsなので、「化学計量学」 by Yuta

ケモメトリックスの適用研究分野

In Silico Data
Miracles by the KY-methods

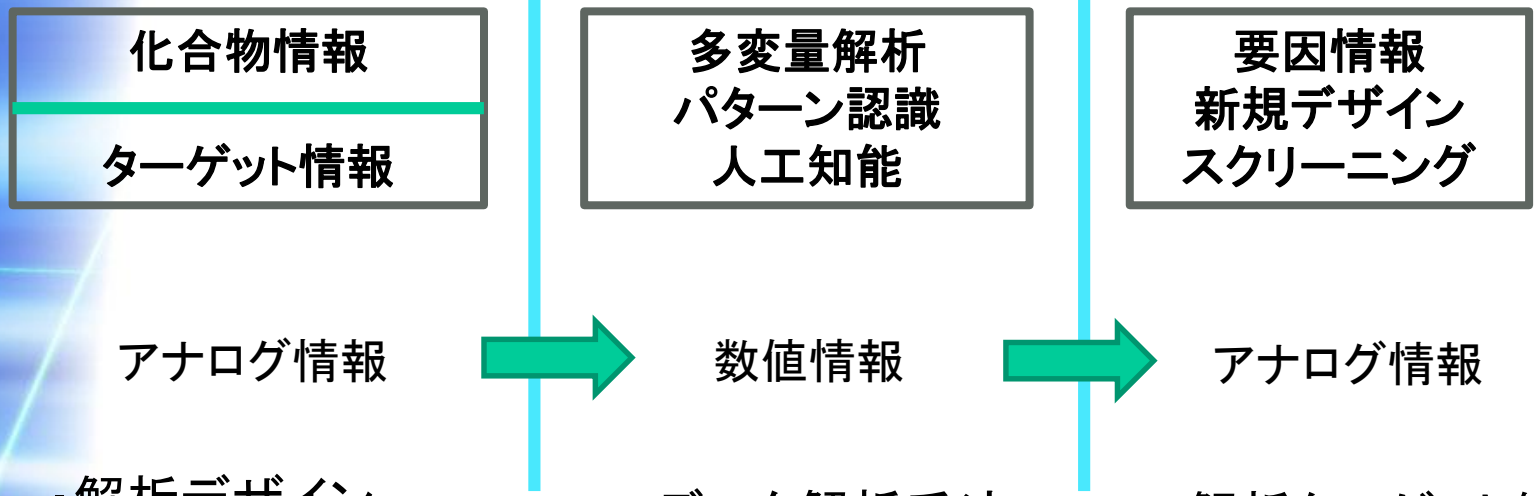
多種多様な分野での研究の基本

（ケモメトリックスは様々な分野で適用できる汎用ツールである）

構造－活性相関（QSAR）
構造－毒性相関（QSTR）
構造－物性相関（QSPR）
メタボロミクス
インシリコ（薬理活性/毒性/物性/ADME）スクリーニング
オフターゲット（off-target）創薬
マルチターゲット創薬
ドラグリポジショニング
テーラーメードモデリング
並列創薬
インシリコンビ
医療解析
バイオ関連解析
その他

□ケモメトリックスによる解析実施の流れ

ケモメトリックス適用基本原理



- 解析デザイン
- データ収集
- プロトコル検討
- その他

- データ解析手法
- パラメーター
- 種々留意事項
- その他

- 解析ターゲット知識
- データの要因解析
- 外挿関連情報／知識
- その他

□ケモメトリックスの化学分野への適用原理

種々データ解析でのケモメトリックス二大適用原理

◇ 宝箱アプローチ

宝箱に相関やメカニズム等の重要な情報が隠されている

＊ 仮設検証型アプローチ（既存アプローチ）
解析前に、重要な情報に関する仮説等が必要。

◇ 発見型アプローチ

様々な重要情報を仮説抜きで発見（探索）する

＊ 仮設検証型アプローチ（既存アプローチ）
解析前に仮設に関する何らかの情報や知識が必要

◇ 宝箱アプローチ：基本原理

多変量解析やパターン認識適用の基本原理

情報等価原理

化合物
蛋白
ゲノム

宝箱
（相関情報）

薬理活性
毒性
副作用
代謝
分解性

情報等価性／必然性

原因情報

説明義務

パターン認識手法

結果情報

◇ 宝箱アプローチ：解析実施基本原理

多変量解析やパターン認識適用の基本原理

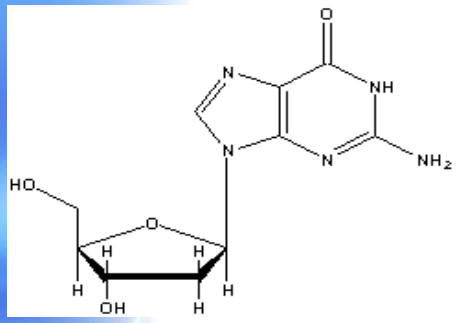
情報等価原理

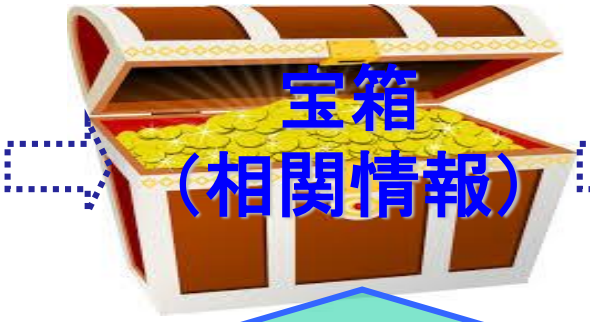
◇発見型アプローチ:

宝箱の中味は
何だろう？

◇ 宝箱アプローチ：化合物関連分野

化合物構造式





薬理活性
毒性
副作用
代謝
分解性

情報の等価性／必然性

種々パラメータ発生

初期パラメータ群
分子量
原子／結合数
HOMO／LUMO
その他

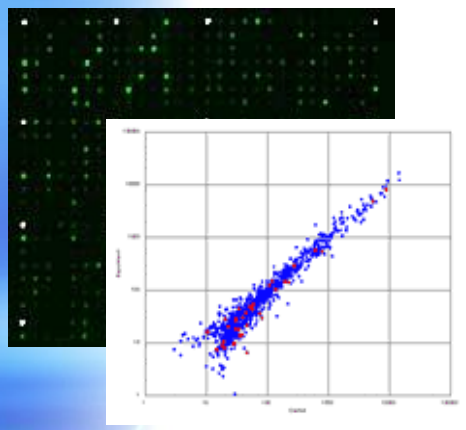
特徴抽出

最終
パラメータ群

種々の解析手法
・判別分析
・クラスタリング
・フィッティング
・その他

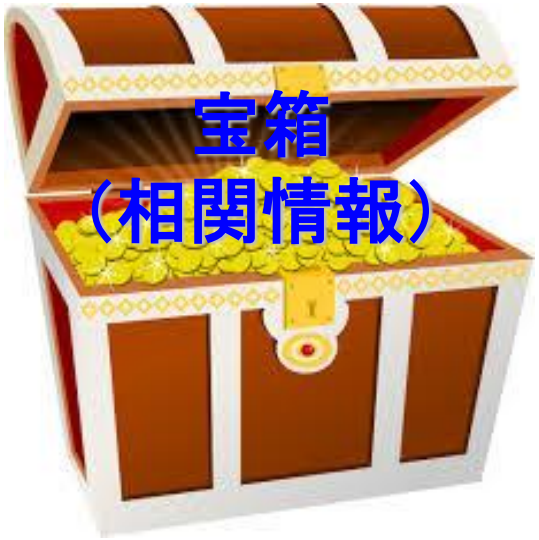
◇ 宝箱アプローチ：バイオ関連分野

発現プロフィールデータ



数値パラメータ

Hs.35092 ;5523
Hs.27935 ; 857
Hs.622 ;3682
Hs.38677 ;2283
・・・・・・



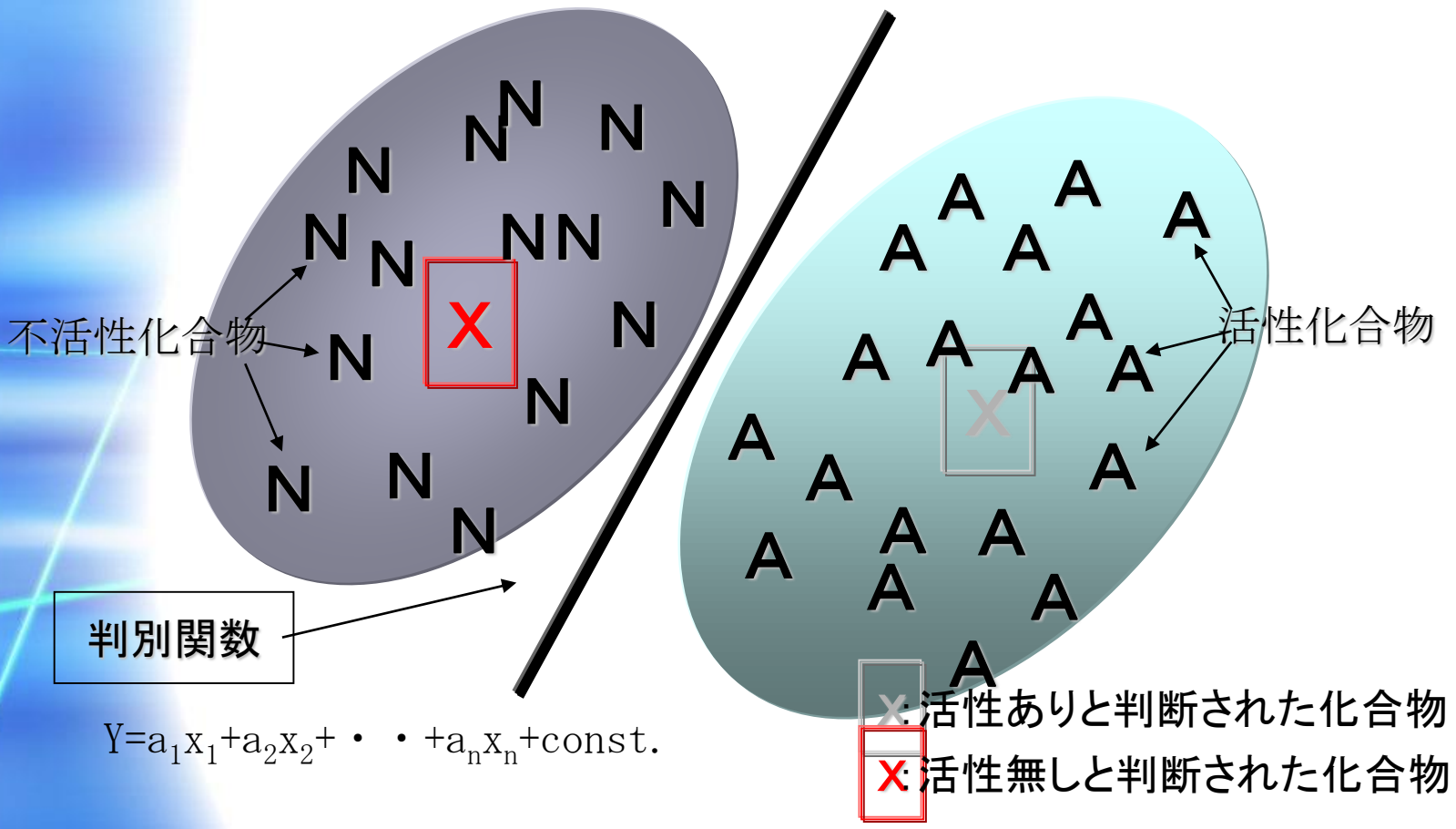
種々の解析手法

判別分析
クラスタリング
マッピング
フィッティング
その他

薬理活性
毒性
副作用
代謝
分解性

二クラス分類の基本概念

判別関数を用いた化合物の分類／予測のイメージ



□ニクラス分類の基本概念

判別関数とドラグデザイン的情報解析

$$Y = a_1x_1 \pm a_2x_2 \pm \cdots \pm a_nx_n \pm \text{const.}$$

Y ： 目的薬理活性／毒性／他

Y ≧ 0	Y ＜ 0
・活性あり ・毒性あり	・活性無し ・毒性無し

□構造－活性／毒性／他との相関解析

係数 $a_i \geqq 0$ の時
パラメータXiの持つ情報は
・活性向上、・毒性強化要因

係数 $a_i < 0$ の時
パラメータXiの持つ情報は
・活性低下、毒性低下要因

構造－活性相関、構造－毒性相関

□ニクラス分類の基本概念

判別関数、重回帰式からの情報取り出し

□パラメータとウェイトベクトルの利用

1. 要因の抽出および明確化：パラメータ利用

最終判別関数や重回帰式中で用いられているパラメータが持つ情報内容の評価

2. 寄与の方向性：ウエイトベクトル利用

①係数 $a_i \geqq 0$ の時 ⇒ パラメータXiの持つ情報は

・活性向上、あるいは毒性強化要因となる

②係数 $a_i < 0$ の時 ⇒ パラメータXiの持つ情報は

・活性低下、毒性低下要因

3. 寄与の相対的な貢献度：ウエイトベクトル利用

係数の絶対値比較による貢献度の評価

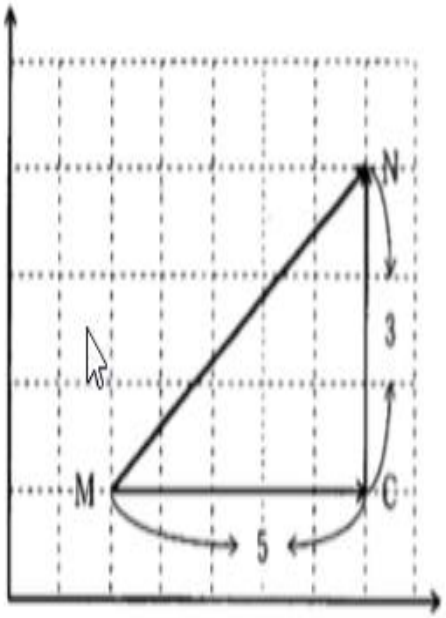
□多変量解析／パターン認識の基本事項

◇サンプル間の**距離**

N次元空間上でのサンプル間の距離の算出は様々な多変量解析／パターン認識手法の基本となる。

サンプル間の距離算出に様々なメトリック手法が展開されている。

右は、ユークリッド距離とシテイブロック距離による距離計算事例である



2次元空間中でA点とB点との距離を計る時、ユークリッド距離では距離D_{MN}を算出し、シテイブロック距離では距離D_{MC}と距離D_{CN}との距離を合わせたものをA点とB点との距離とする。

M＝（2，1）
N＝（7，4）

ユークリッド距離 ＝ $D_{MN} = (5^2 + 3^2)^{1/2} = 5.83$

シテイブロック距離 ＝ $D_{MC} + D_{CN} = 5 + 3 = 8$

図1．　ユークリッド距離とシテイブロック距離

◇サンプル間の距離：**距離基準**

(1) ミンコフスキー (MINKOWSKI)距離

$$D = \left[\sum_{i=1}^{d} (X_{Mi} - X_{Ni})^{k} \right]^{1/k}$$

(2) ユークリッド(EUCLIDEAN) 距離

ユークリッド距離はミンコフスキー距離の式において、kが2の時にあたる。

$$D = \left[\sum_{i=1}^{d} (X_{Mi} - X_{Ni})^{2} \right]^{1/2}$$

(3) シテイブロック(CITY BLOCK)距離

シテイブロック距離は2パターン間の最短距離をとるのではなく、直交する2線の距離の総和をとるものである。

$$D = \sum_{i=1}^{d} \left[X_{Mi} - X_{Ni} \right]$$

(4) キャンベラ (CANBERA)距離

$$D = \frac{\sum_{i=1}^{d} \left[X_{Mi} - X_{Ni} \right]}{\sum_{i=1}^{d} \left[X_{Mi} + X_{Ni} \right]}$$

(5) ハミング(HAMMING) 距離

ハミング距離は1／0のバイナリデータで利用される事が多いが、ORとANDの識別を効率良く行う事が出来る。

$$D = \sum_{i=1}^{d} \left[X_{Mi} + X_{Ni} - 2X_{Mi}X_{Ni} \right]$$

尚、1／0のバイナリーデータを用いた時、ハミング距離とシテイブロック距離は同じものとなる。

(6) 谷本距離

谷本距離はハミング距離がその性格上、1が少ないデータは不利に評価されるという欠点を改良したものである。

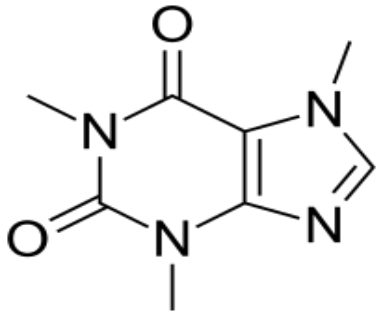
$$D = \frac{\sum_{i=1}^{d} \left[X_{Mi} + X_{Ni} - 2X_{Mi}X_{Ni} \right]}{\sum_{i=1}^{d} \left[X_{Mi} + X_{Ni} - X_{Mi}X_{Ni} \right]}$$

□化学データ（パラメーター）の種類

化合物に起因するデータを「化学データ（パラメーター）」とする
構造式の有無や内容により、以下の3種類に分類される

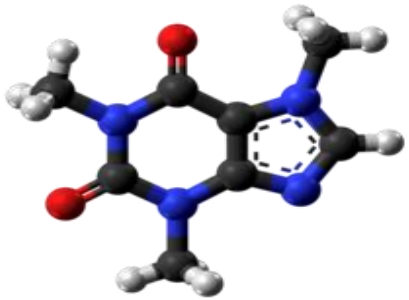
構造式不要

物性データ
機器スペクトルデータ
医療関連データ



2次元構造式関連データ

分子式関連パラメーター
トポロジカルパラメーター
部分構造パラメーター
フラグメントパラメーター



3次元構造式関連データ

トポグラフィカルパラメーター
3次元構造関連データ
力場関連パラメーター
量子化学関連パラメーター

□化学データ(パラメーター)の種類

◇化合物に由来する化学データ （パラメーター）

化合物関連データ解析を行うにはパラメーター、即ち数値データが必要となる

化合物構造式由来のパラメーター

■ トポロジカルデータ
分子構造インデックス：原子数（原子種）、結合数（結合種）、リング数、その他
様々なインデックス値：ＨＯＳＯＹＡインデックス、分子結合インデックスＭＣ値
パス値インデックス、

■ トポグラフィカルデータ
化合物の３次元的形状に関するパラメータ
化合物全体構造 ：ボックスパラメータ、対称パラメータ、
立体格子パラメータ、その他
化合物部分構造 ：ステリモルパラメータ、

■ 物理化学データ
分子に関する様々な物性データ ： 分子屈折率、分子量、ＬＯＧＰ、融点、沸点
分子容積、分子表面積、その他
分子軌道法より得られる様々なパラメータ：電子密度、ＨＯＭＯ、ＬＵＭＯ、他
分子力学計算から得られるパラメータ： 種々歪みエネルギー
種々スペクトルより得られるデータ： 種々スペクトルデータ

■ その他のデータ
部分構造パラメータ： 部分構造の有無、部分構造数、
部分構造単位の様々なパラメータ値計算、
演算パラメータ１ ： 記述子間の演算により得られるパラメータ（＋－ｘ÷ Log）
演算パラメータ２ ： 他の解析手法より算出されたパラメータ
ダミーパラメータ ： 有るパターン存在の有無（１／０）に関するパラメータ

機器スペクトル由来のパラメーター

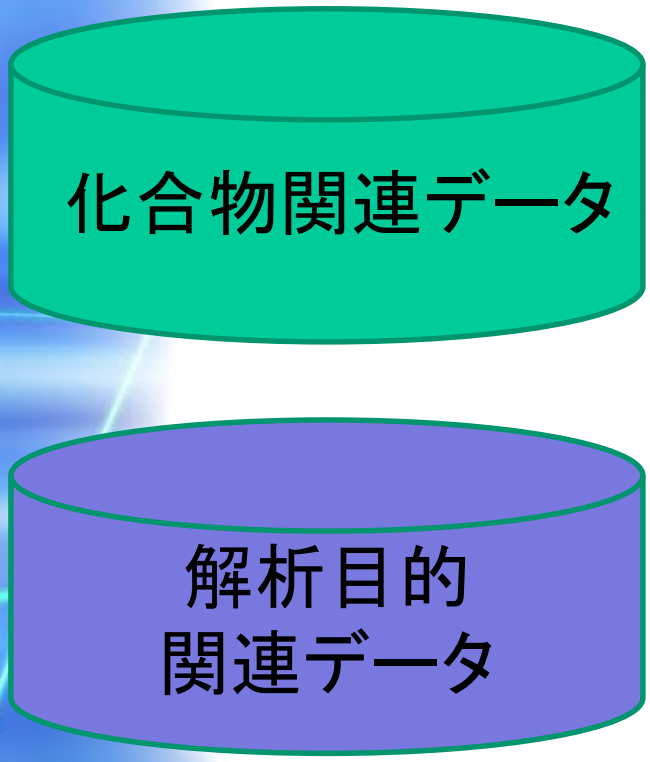
Mass
IR
H-NMR
C-NMR
UV
GC
HPLC
Raman
X線分析
その他

化合物に起因する解析目的パラメーター

薬理活性
毒性
ADME
物性
環境毒性

□化学データ解析での必要データ

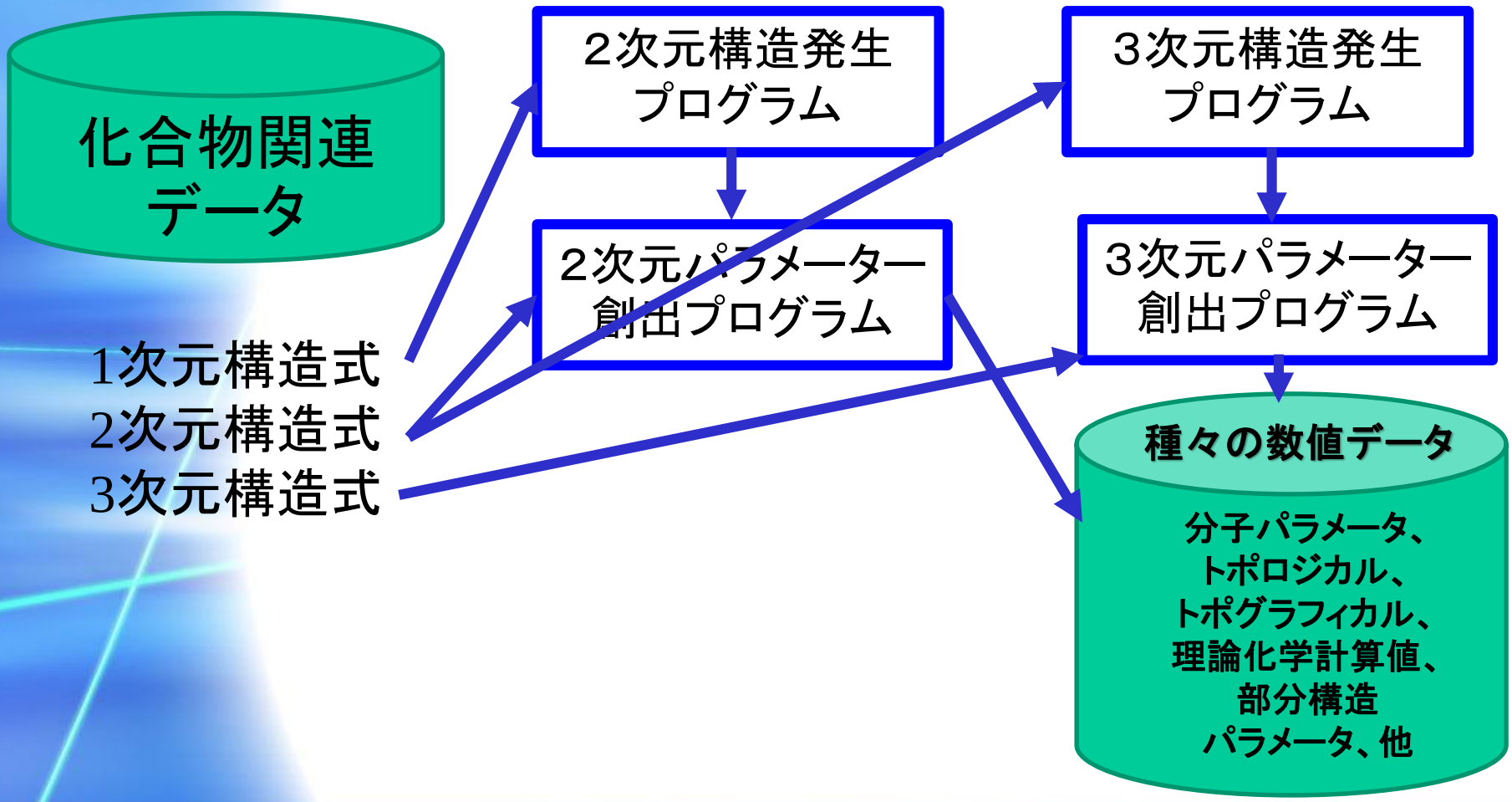
◇汎用的なデータ解析の流れ図：入力データ関連



＊化合物構造式（1／2次元）
＊物性データ
＊機器スペクトルデータ
＊バイオ関連データ
＊医療関連データ（画像等）

＊薬理活性
＊毒性関連データ
＊機器スペクトルデータ
＊バイオ関連データ
＊患者関連データ

◇化合物構造式を用いたパラメーター発生



◇汎用的なデータ解析の流れ図：データ前処理関連

解析目的と無関係な情報を有するパラメーター除去（特徴抽出）

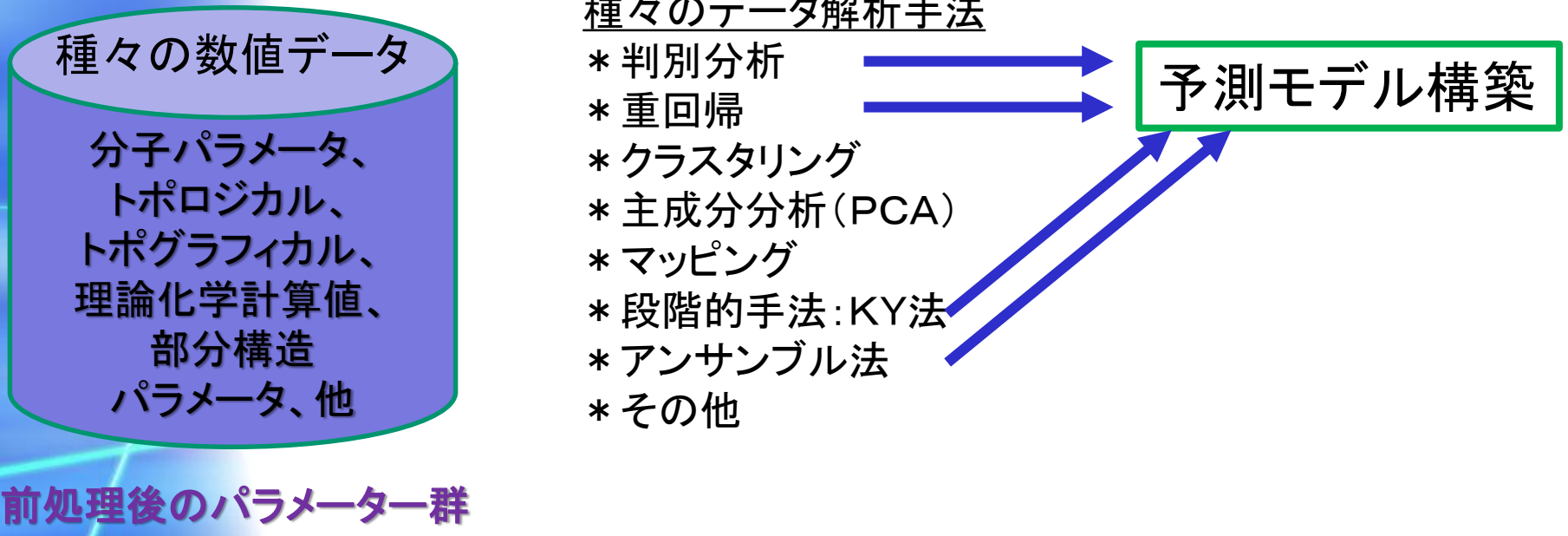
種々の数値データ

分子パラメータ、
トポロジカル、
トポグラフィカル、
理論化学計算値、
部分構造パラメータ、
演算パラメーター
その他

パラメーター選択（特徴抽出）の実施

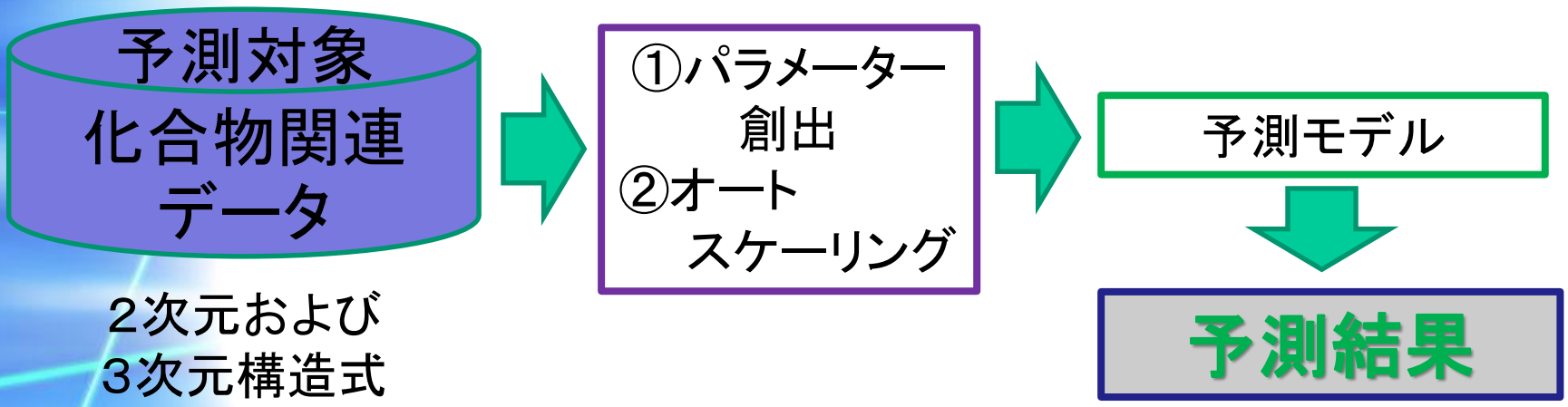
＊欠陥データ処理：
　欠損データ、0値データ、同値データ、他
＊統計的処理：
　単相関、重相関、Fisher比、他
＊データ解析手法による個別特徴抽出
　主成分解析、パーセプトロン、遺伝的アルゴリズム、他
＊パラメーター桁の調整
　オートスケーリング

◇汎用的なデータ解析の流れ図：多変量解析／パターン認識関連

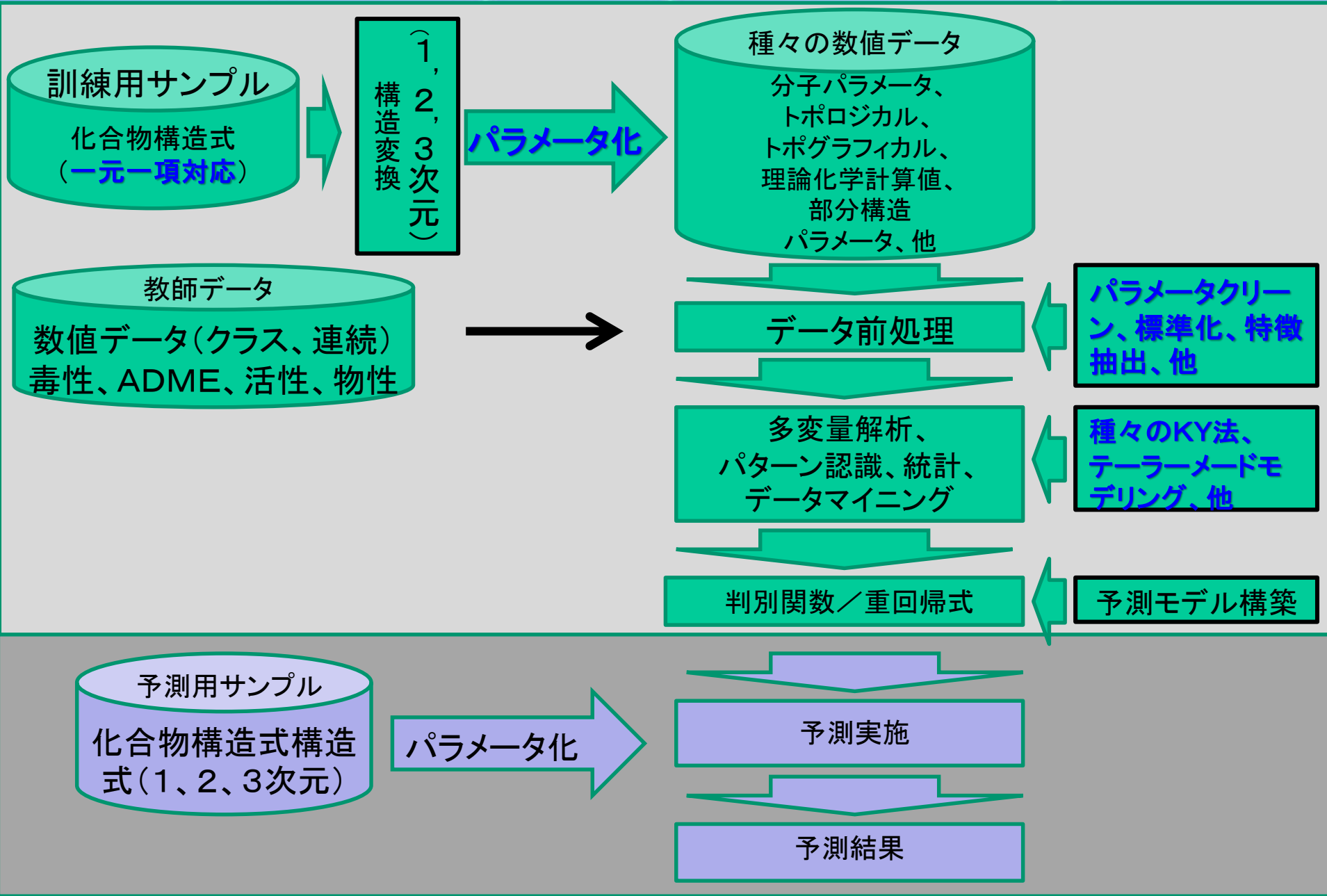


◇汎用的なデータ解析の流れ図：予測の実施

＊予測モデルを構成するパラメーターを化合物構造式から創出
＊構造式から発生できないパラメーターは外部パラメーターとして導入
＊オートスケーリングを実施



◇汎用的なデータ解析の流れ図：解析全体の流れ



データ解析手法の
簡単な紹介

□ データ解析手法の機能的分類

統計関連情報の扱い方によりデータ解析手法は大きく二種類に分類される

1. パラメトリック手法

解析に用いるパラメーターの母集団分布情報を用いて解析する手法
分布（正規分布等）等の情報、平均値、分散、サンプル数は大きいものが良い
統計の検定等を実施するときに利用される

2. ノンパラメトリック手法

解析に用いるパラメーターの母集団分布情報が不明な環境で解析する手法
分布型は問わない、サンプル数も小さくても実施可能、他
制限事項が少ないので、広範囲にわたって適用できる
基本的に多変量解析／パターン認識として展開される手法はノンパラメトリック

◇ データ解析手法の機能的分類

解析アプローチの差異に伴い大きく以下の解析手法に分類される
個々の手法での代表的な手法をリストする

1. 判別分析：二クラス分類、多クラス分類
2. 重回帰（フィッテイング）：単回帰、重回帰
3. クラスター分析：階層型クラスタリング、非階層型クラスタリング
4. 種々マッピング：主成分分析（PCA）、非線形写像（NLM）
5. チャート表示：手法⇒レーダーチャート、顔チャート、ラインチャート、
6. 決定木：手法⇒C5.0、CART
7. アンサンブル学習法：手法⇒AdaBoost、ランダムフォレスト
8. 段階的手法：手法⇒KY法
9. 最適化手法：最小二乗法、シンプレクス法、遺伝的アルゴリズム

◇個別手法の特徴と適用内容：クラスター分析

ModelBuilderの画面より

◇個別手法の特徴と適用内容：クラスター分析
クロスた一分析は以下の基準にて様々な手法に分類される

クラスター化アプローチ

■階層的クラスタリング
解析結果はデンドログラムとして出力される

■非階層的クラスタリング
解析結果は単にクラスターの数とメンバー

クラスター化アルゴリズム

①Division Method
（分割法）
② Aggregative Method
（凝集法）

融合法の種類

① 最近隣法 (NEAREST NEIGHBOR METHOD)
② 最遠隣法 (FURTHEST NEIGHBOR METHOD)
③ 群平均法 (GROUP-AVERAGE METHOD)
④ 重心法 (CENTROID METHOD)
⑤ メジアン法 (MEDIAN METHOD)
⑥ ワード法 (WARD METHOD)
⑦ 可変法 (FLEXIBLE METHOD)

◇個別手法の特徴と適用内容：クラスター分析
クラスタリングの融合手法

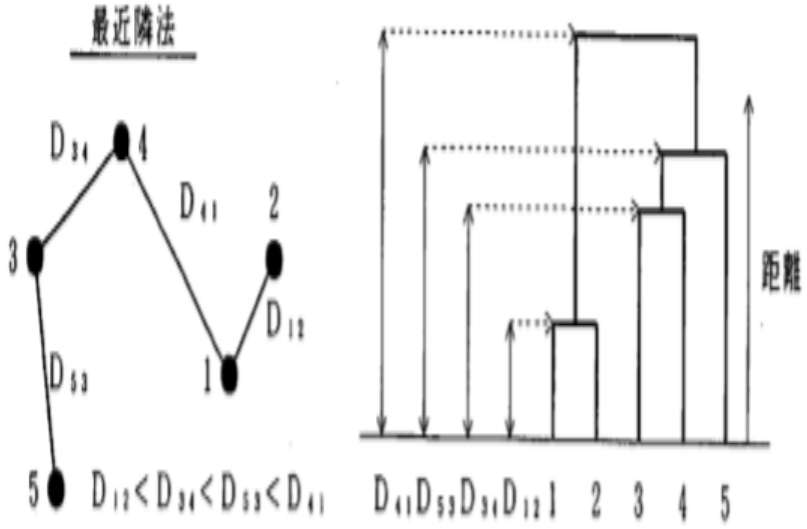


図2、 最近隣法によるクラスタリング手続きとデンドログラム

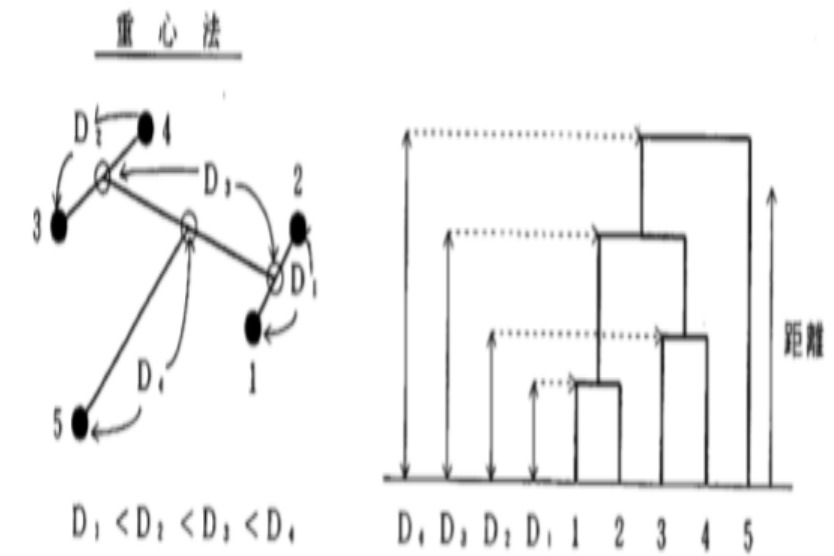
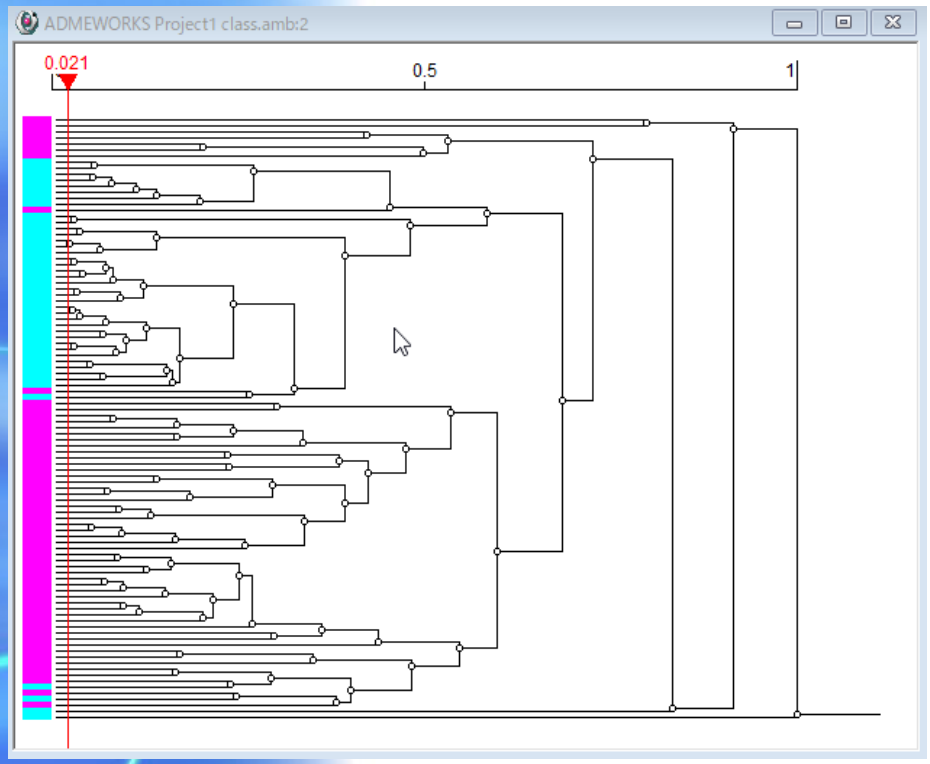


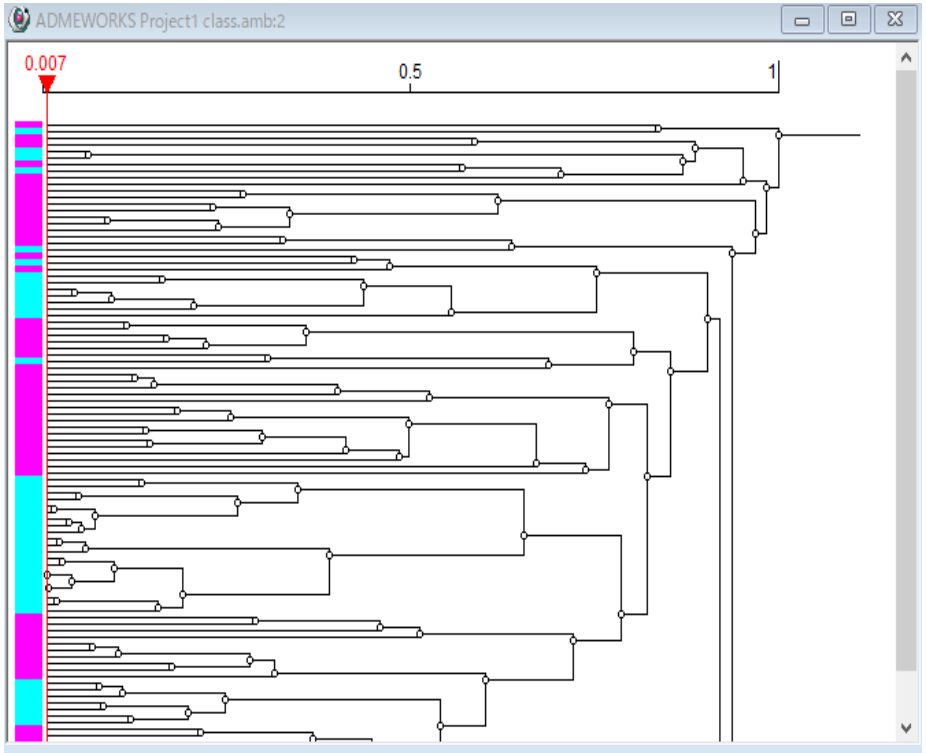
図3．重心法によるクラスタリング手続きとデンドログラム

ModelBuilderの画面より

◇個別手法の特徴と適用内容：クラスター分析

＊クラスタリングはサンプル同士の相対的な位置関係や近隣関係を俯瞰して見ることが出来る
＊メトリックや融合手法の差異によりクラスタリング結果が大きく変化するので要因解析は注意





ModelBuilderの画面より

◇個別手法の特徴と適用内容：マッピング

■主成分分析（PCA）
・データ解析的にはリニアプロジェクション手法
・サンプル空間中のサンプルの位置関係を変えない
・オリジナルのサンプル空間を俯瞰する方向性を変えて空間を見直す手法
・PCA適用前と後とで次元（パラメーター）数の変化はない

■非線形写像（NLM）
・データ解析的にはマッピング（写像）手法
・サンプルを人間が可視できる二次元/三次元上に分散
・サンプル空間を強制的に2/3次元に圧縮する
・非線形写像実施後は、サンプル空間の次元は2/3次元となる

◇個別手法の特徴と適用内容：Non　Linear　Mapping（NLM）

＊NLMにより、最初のN次元空間が
可視できる2次元空間に変換された。

＊NLMでは、元のN次元空間における
サンプル間の位置関係を保ちつつ
新しい2次元空間上に再配置される。

＊2次元空間上の第6サンプルは、
その他のサンプル（1～5、7～9）との
相互位置関係（空間上の距離）をN次元
空間上における関係と略同じとなる

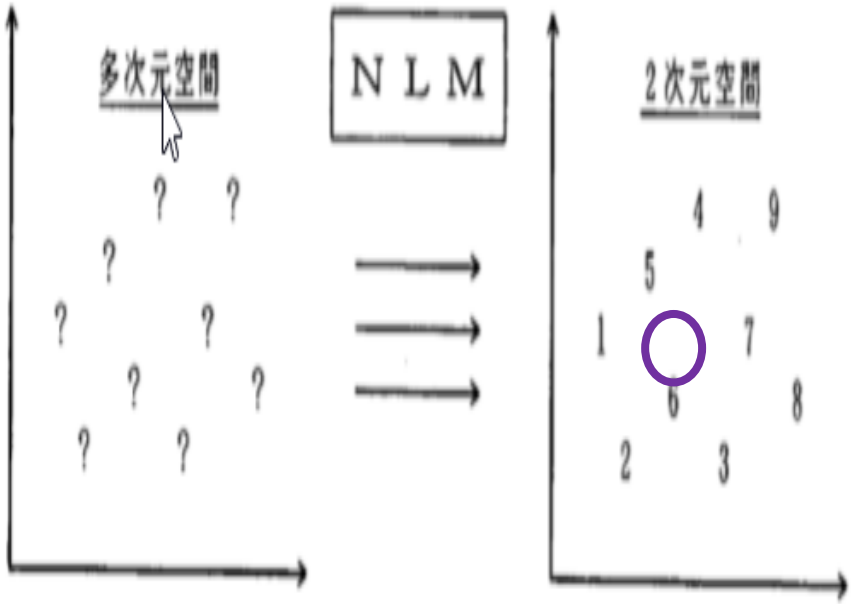


図5－6　ノンリニアマッピングによる多次元空間の2次元空間への写像

◇個別手法の特徴と適用内容：Non　Linear　Mapping（NLM）

多次元（d次元）空間上のパターンA，B

$$PA = (X_{A1},\ X_{A2},\ X_{A3},\ \cdots\cdots,\ X_{A,d-1},\ X_{A,d})$$

$$PB = (X_{B1},\ X_{B2},\ X_{B3},\ \cdots\cdots,\ X_{B,d-1},\ X_{B,d})$$

2次元空間上のパターンM，N

$$P2A = (Y_{A1},\ Y_{A2})$$

$$P2B = (Y_{B1},\ Y_{B2})$$

$$DX_{AB} = \sum_{i=1}^{d} \{(X_{Ai} - X_{Bi})^2\}^{1/2} \quad \text{————} \quad (\ )$$

$$D2_{AB} = \sum_{i=1}^{2} \{(Y_{Ai} - Y_{Bi})^2\}^{1/2} \quad \text{————} \quad (\ )$$

$$Ei = \frac{1}{\sum_{A\neq B}^{n} DX_{AB}} \sum_{A\neq B}^{n} \frac{(DX_{AB} - D2_{AB})^2}{DX_{AB}} \qquad (i = 1,\ 2,\ \sim,\ n)$$

ここで、nはパターンの数を示す。

このエラー関数の極小化は（Y_{i1}，Y_{i2}）（i＝1～n）の初期チャートを適当に与え

$$\frac{\partial E(Y_{ij})}{\partial Y_{ij}} = 0 \quad (i = 1 \sim n;\ \ j = 1,\ 2) \quad \text{————} \quad (\ )$$

$$Y_{ij}(m+1) = Y_{ij}(m) - k\Delta_{ij}(m) \quad \text{————} \quad (\ )$$

ここで$\Delta_{ij}(m) = \dfrac{\partial E(m)}{\partial Y_{ij}(m)} \Big/ \left| \dfrac{\partial^2 E(m)}{\partial Y_{ij}(m)^2} \right|$　で、k≒0.3～0.4

◇個別手法の特徴と適用内容：

ニューラルネットワークによる次元圧縮

＊入力層のN個のパラメーターから中間層のユニット数（図は3）に次元を変換する手法

＊NLMと異なり、サンプル間の相互位置関係はキープされない

＊新たな次元は、ニューラルネットワークが解析目標とするクラスデータ等を説明する情報を含んでいる

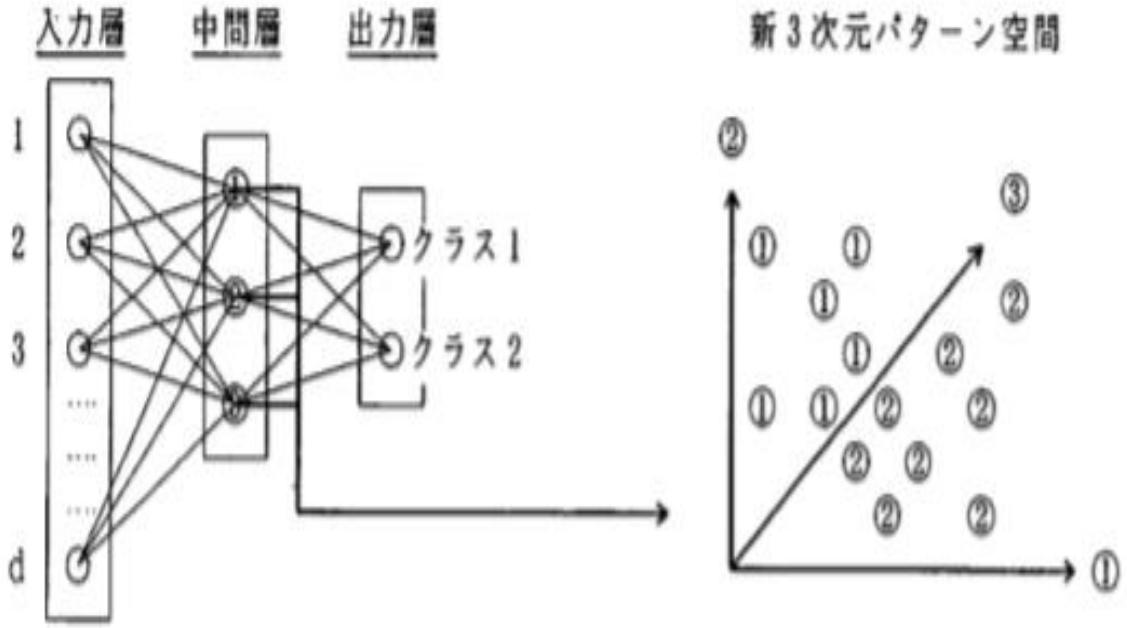


図５－７．バックプロパゲーションによる写像。ｄ次元⇨３次元
（新たな次元は各パターンがクラス１とクラス２とに分類されやすいように３次元空間上に分布していることに注意）

アンサンブル学習法

AdaBoostやランダムフォレスト等で採用されている解析手法
特徴：
弱分類機を条件を変えつつ多数用い、最終的な分類結果は個々の分類結果データを統合して最終結果とする。

従来の分類手法を組み合わせて使う「メタ解析手法」である。
＊「KY法」もメタ解析手法となる

◇ニューラルネットワーク

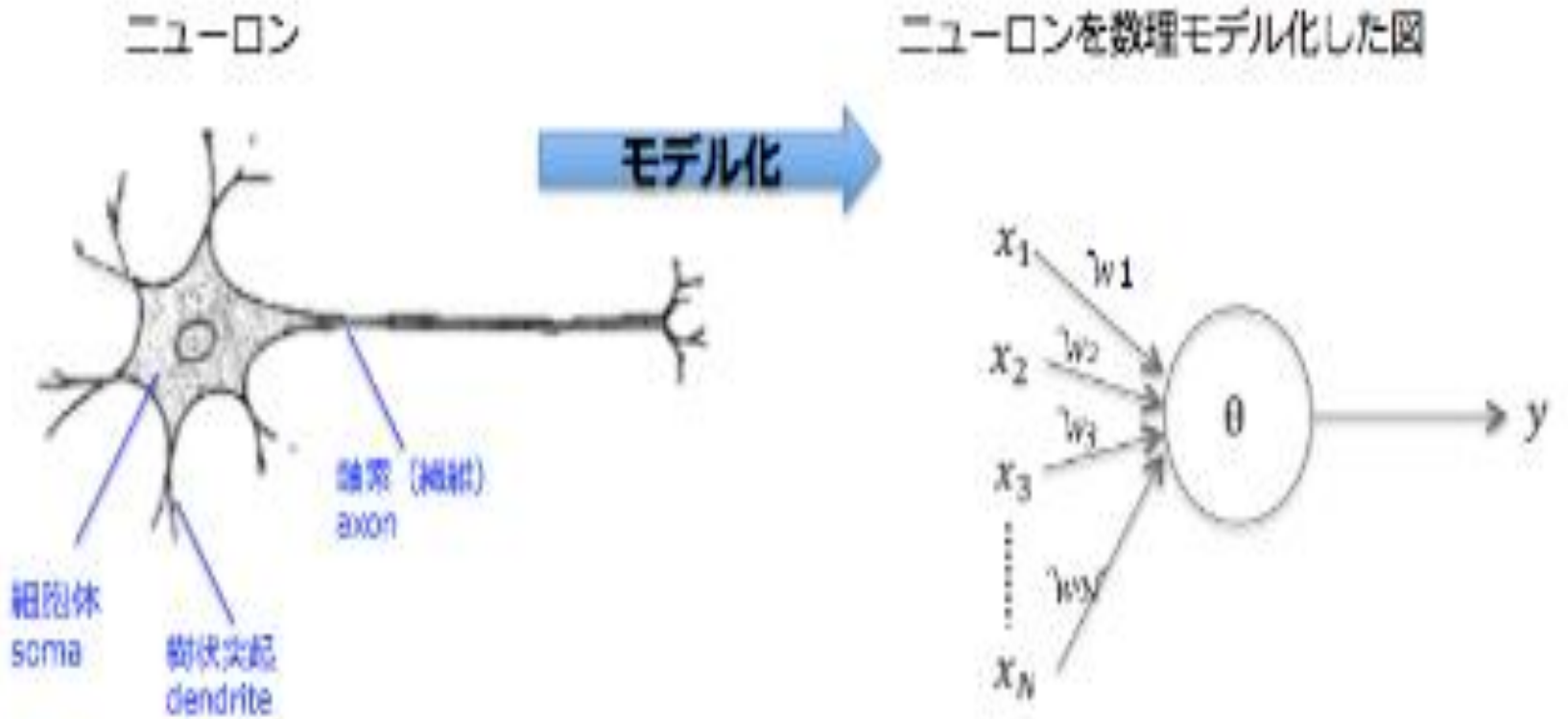
＊ニューラルネットワークは単純なネットワーク構造を利用したパーセプトロンを
基本として発展
＊パーセプトロンは簡単な二分類問題も解決できないということで、衰退
＊パーセプトロンの限界を打破したアプローチとしてニューラルネットワークが提案

パーセプトロン：線形分類機

ニューラルネットワーク：非線形分類機

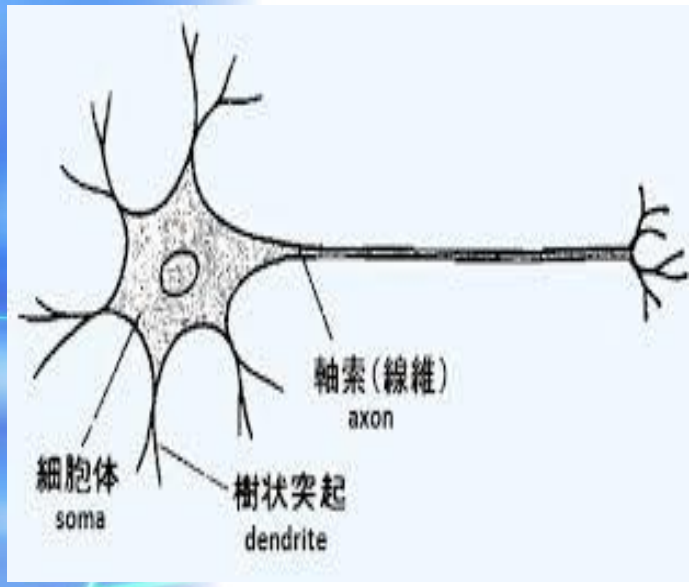
＊パーセプトロンは脳の機能を模したアプローチとして開発され、最終目標は
AI(人工知能）への展開であった
＊ニューラルネットワークはその改良系で、ネットワーク構造が多層構造となった
＊最近の深層学習はニューラルネットワークのネットワーク構造をさらに複雑にした

ロパーセプトロン

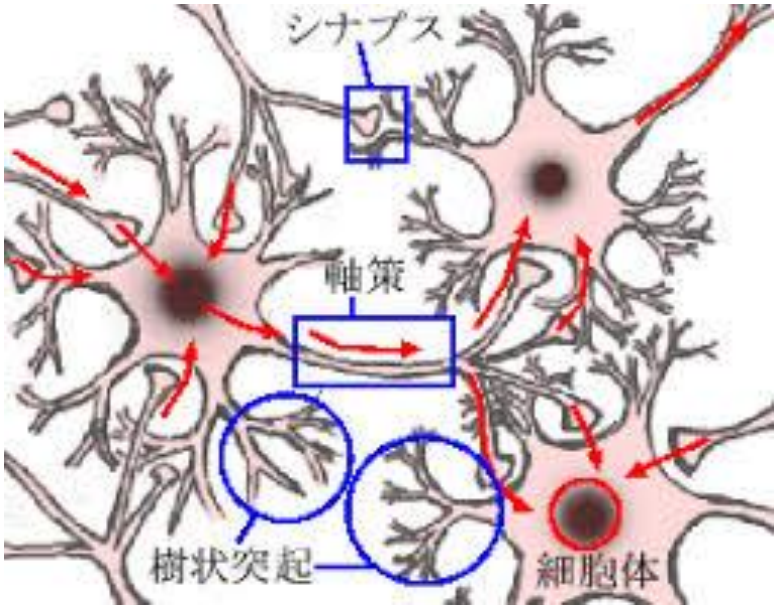


https://udemy.benesse.co.jp/ai/neural-network.html

単ニューロンモデル



ネットワークニューロンモデル



http://www.tamagawa.ac.jp/teachers/aihara/kouzou.html

http://www.sys.ci.ritsumei.ac.jp/project/theory/nn/nn.html

パーセプトロン

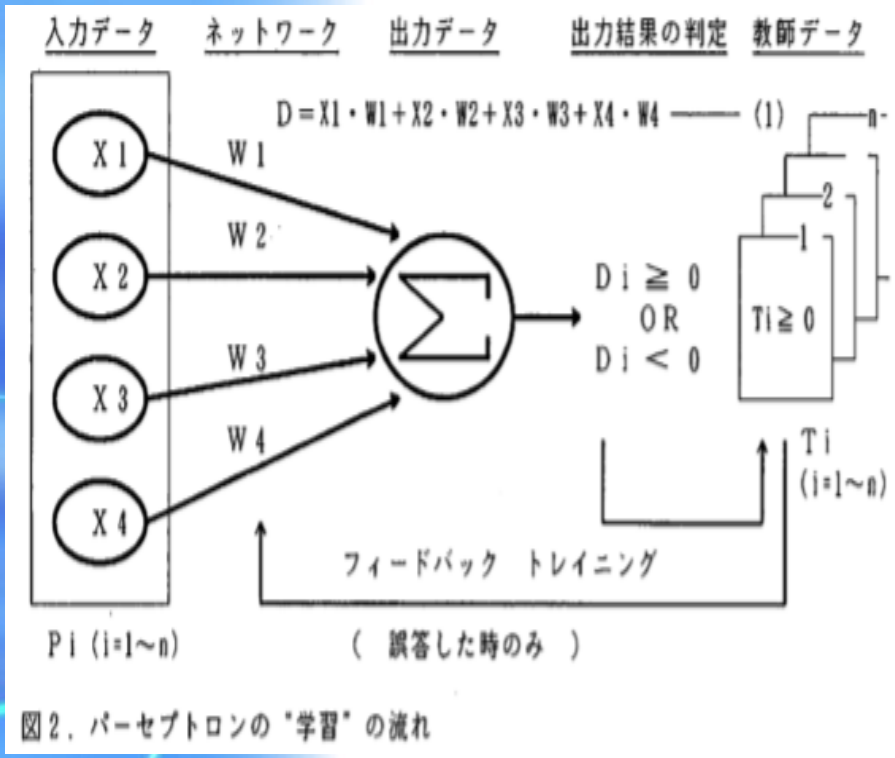
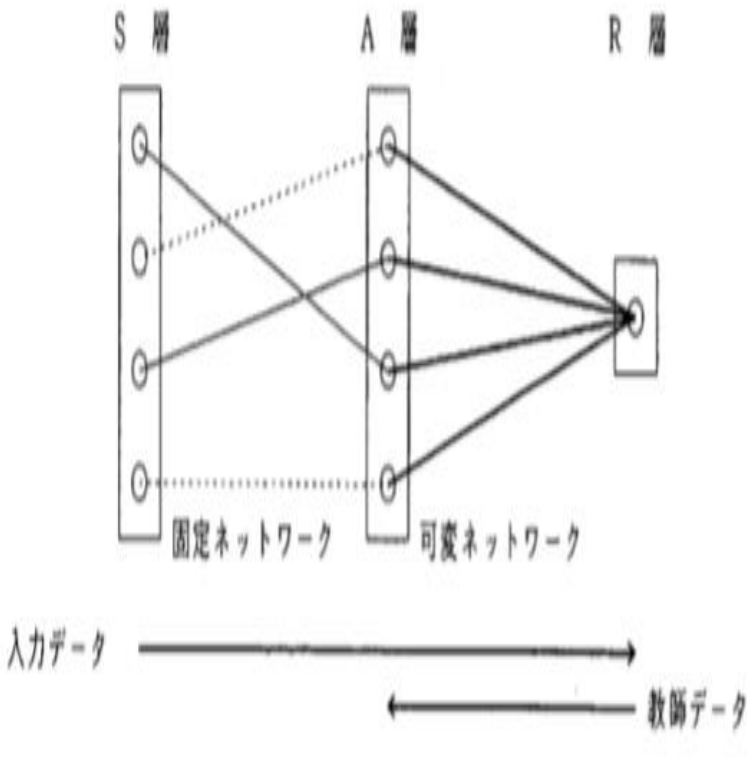


図2．パーセプトロンの"学習"の流れ

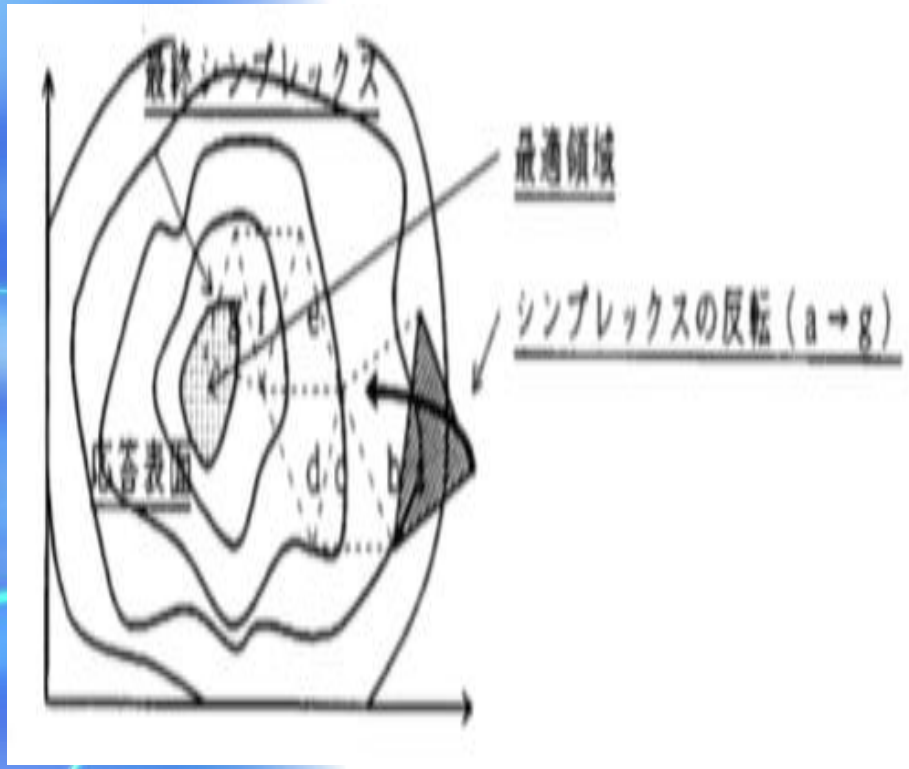
単ニューロンモデル

ニューラルネットワーク



ネットワークニューロンモデル

◇最適化手法：シンプレックス法



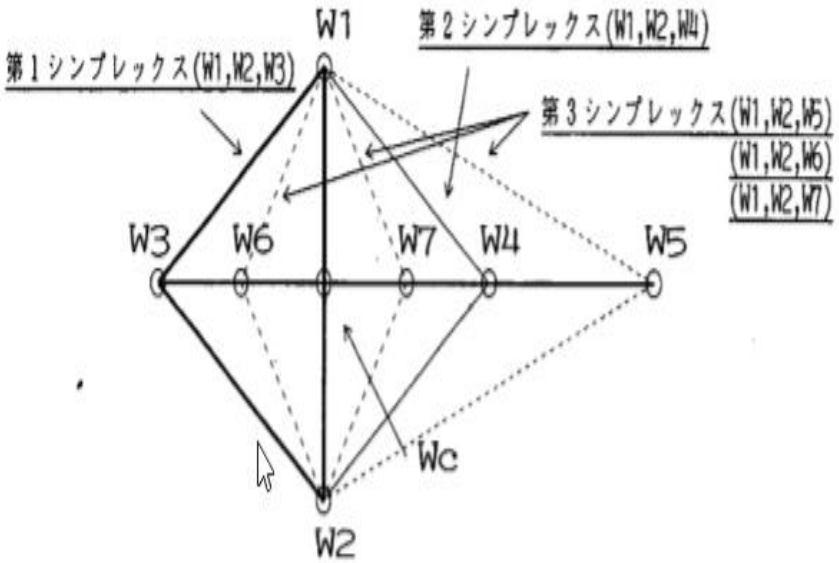


図6－2　シンプレックス反転に関するルール

◇最適化手法：遺伝的アルゴリズム

遺伝子の分裂／増殖／突然変異等の動きをシミュレーションするアプローチ

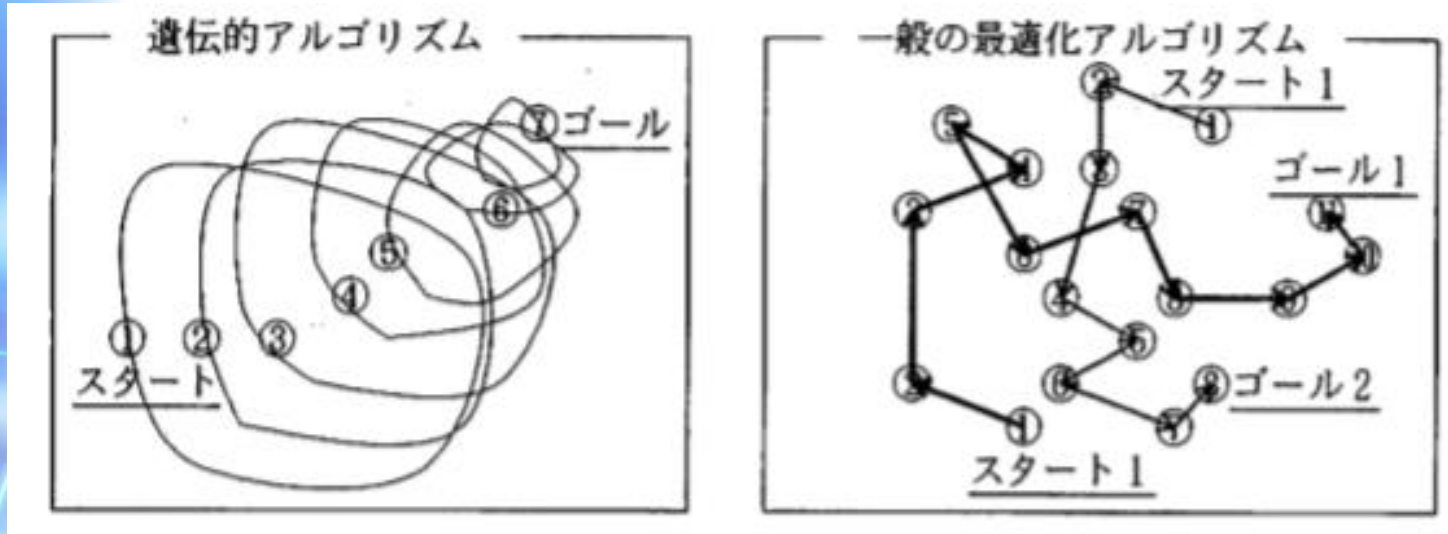


図3．遺伝的アルゴリズムと一般の最適化アルゴリズムによる最適領域の探索過程

1．増殖（MULTIPLICATION）
2．交叉（CROSSING-OVER）
3．突然変異（MUTATION）
4．淘汰／選択（SELECTION）

◇最適化手法：遺伝的アルゴリズム

1. 増殖 (MULTIPLICATION)

2. 交叉 (CROSSING-OVER)
情報の取替により、情報の変換を果たすものである。

1011010010111011001001011 遺伝子A

1001110011000011001001011 新遺伝子BA

1011010010111010111010100011 新遺伝子AB

1001110011000010111010100011 遺伝子B

異なる遺伝子パターンをもつ2本の遺伝子がある時、これら2本の遺伝子を交叉させる。 この結果、新たに2本の遺伝子が誕生するがこれらの遺伝子は親となるAおよびBの遺伝子の形質を交叉させた点を中心としてそれぞれあわせ持つことがわかる。

3. 突然変異 (MUTATION)
この突然変異では遺伝子の持つ情報がある部位で変化して対立遺伝子となる。

1011010010110110010010111 遺伝子A

↓ 突然変異

1011010010100110010010111 遺伝子A'

4. 淘汰／選択 (SELECTION)
様々な変換パターンにより構築された遺伝子が淘汰により悪い形質（問題解決に取って望ましくない）を持つものが取り除かれてゆくことである。 生物学的にはまわりの環境に適合した形質を持つ固体だけが生き残り（選択され）、次世代へと情報（形質）を伝えてゆくことを意味する。

淘汰

101 …… 111 遺伝子A → ×
010 …… 110 遺伝子B → 010 …… 110 遺伝子B
001 …… 001 遺伝子C → ×
…… → …… ……
100 …… 011 遺伝子Z → ×

◇最適化手法：ニューラルネットワーク

閉じたネットワーク構造を有するニューラルネットワーク

ボルツマンマシンおよびホプフィールドネット

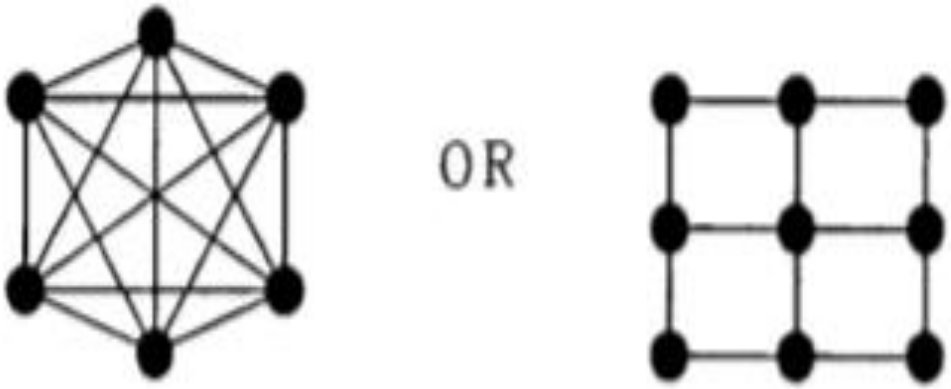


図1．ボルツマンマシン及びホップフィールドネットのネットワーク構造

◇KY法について

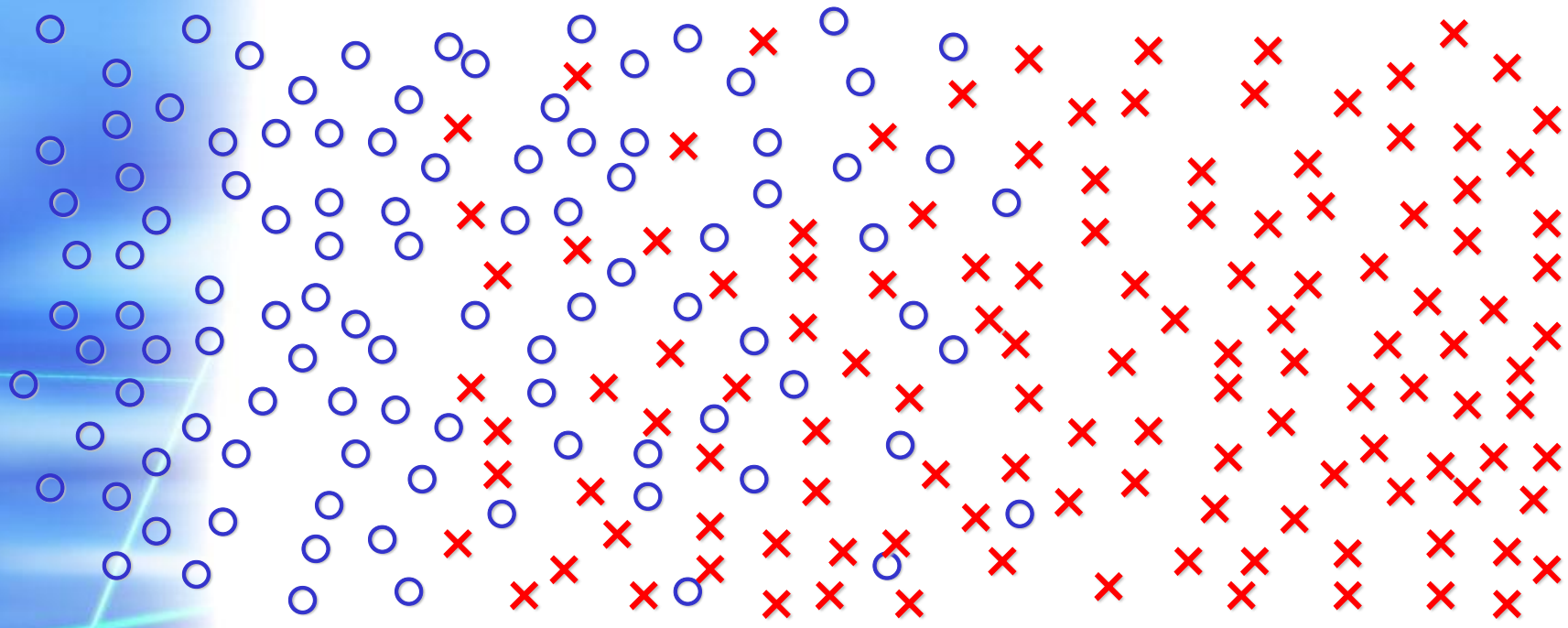
KY法(K-step Yard sampling methods)の特徴:以下の二大特徴を持つとKY法となる

①多段階解析手法:一回の解析でなく、サンプル群を変えて何回も解析する
②メタ解析手法;KY法の各ステップで使われるデータ解析手法は従来法である

①の特徴により、各段階においてサンプル群のリサンプリングが実施される。
繰り返しデータ解析が実施される点でアンサンブル法に似ているが、
・アンサンブル法、適用対象となるサンプル母集団は変化しない
・KY法では、サンプル母集団が変化し、これに対してデータ解析が実施される

②メタ解析手法という点でアンサンブル法と同じであるが、以下の点で大きく異なる
・KY法ではステップ単位で適用されるデータ解析手法を変えることが出来る
・アンサンブル法では、基本的に適用されるデータ解析手法は同じ手法を用いる

◆ 一般的なサンプル空間



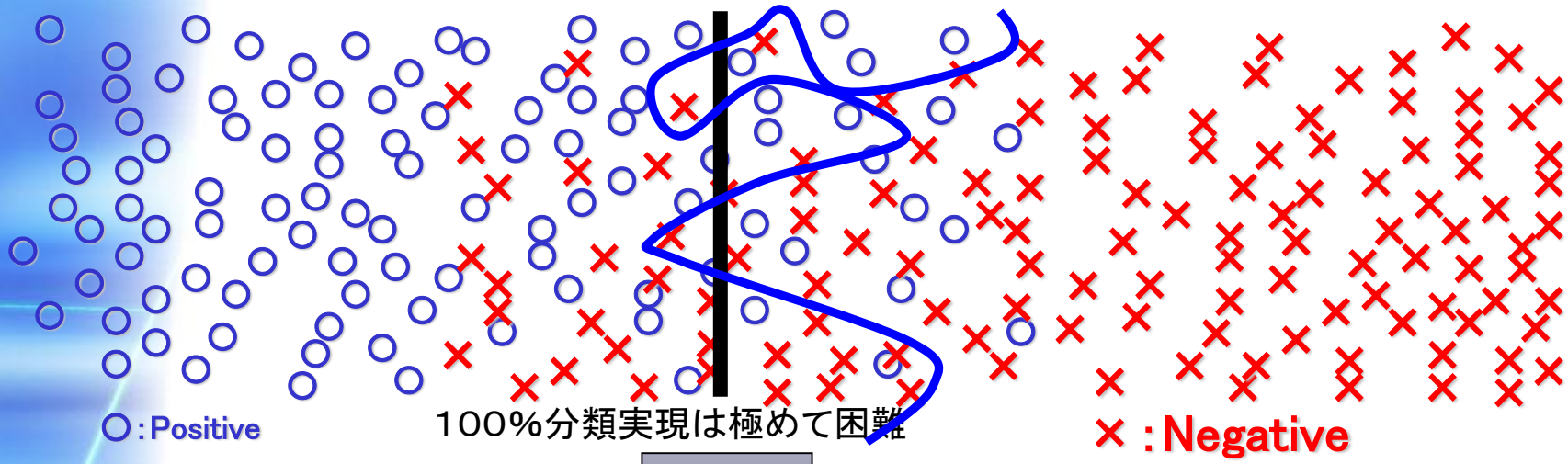
○：Positive

×：Negative

100%分類の実現は極めて困難

◆パターン分布とパターン分類：識別不可能な場合

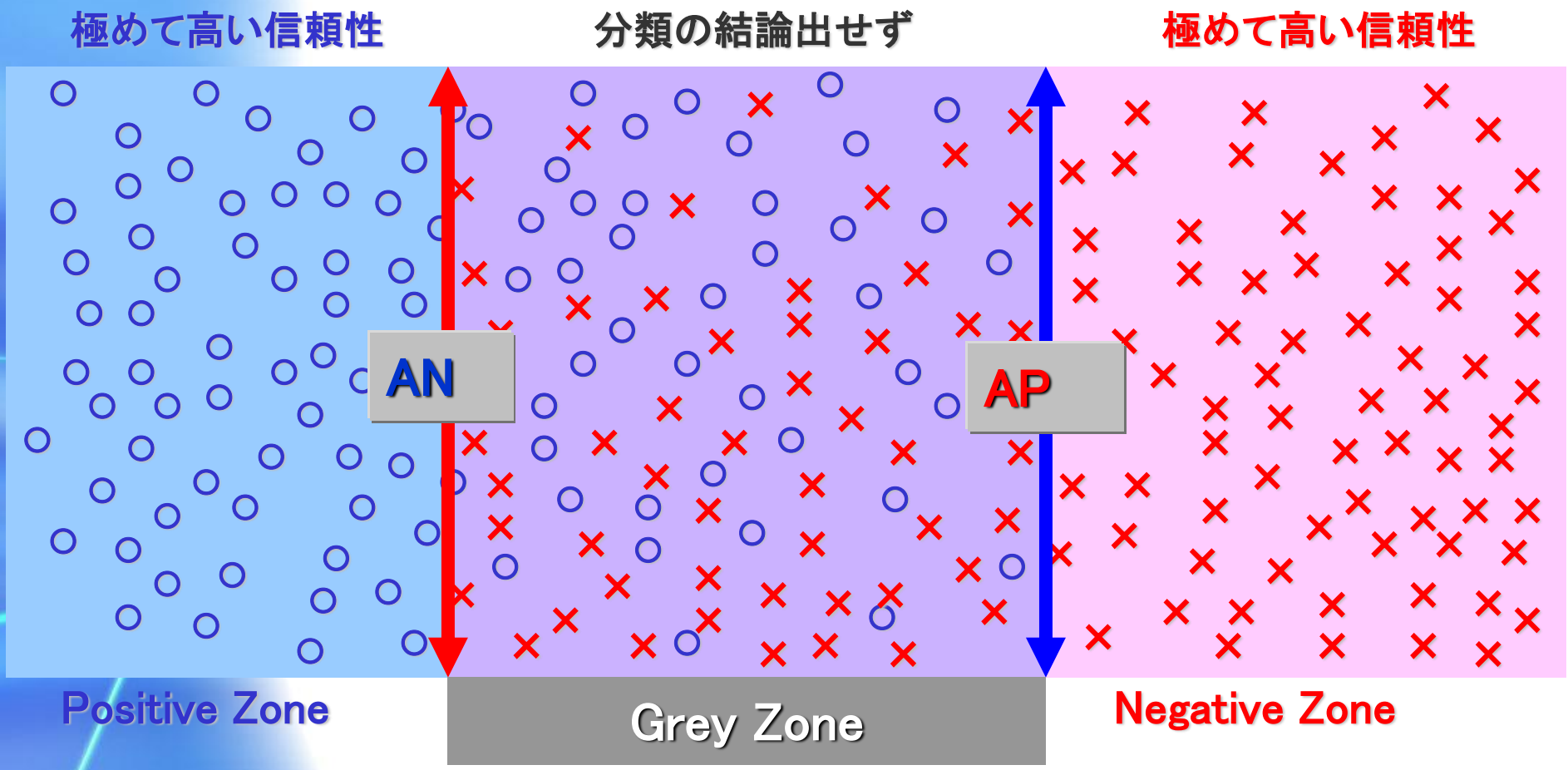
100%分類不可能な識別線（判別関数）



新たなサンプル空間を創出する、非線形化する、
次元の拡張等行ったとしても100%分類の可能性は殆ど無い

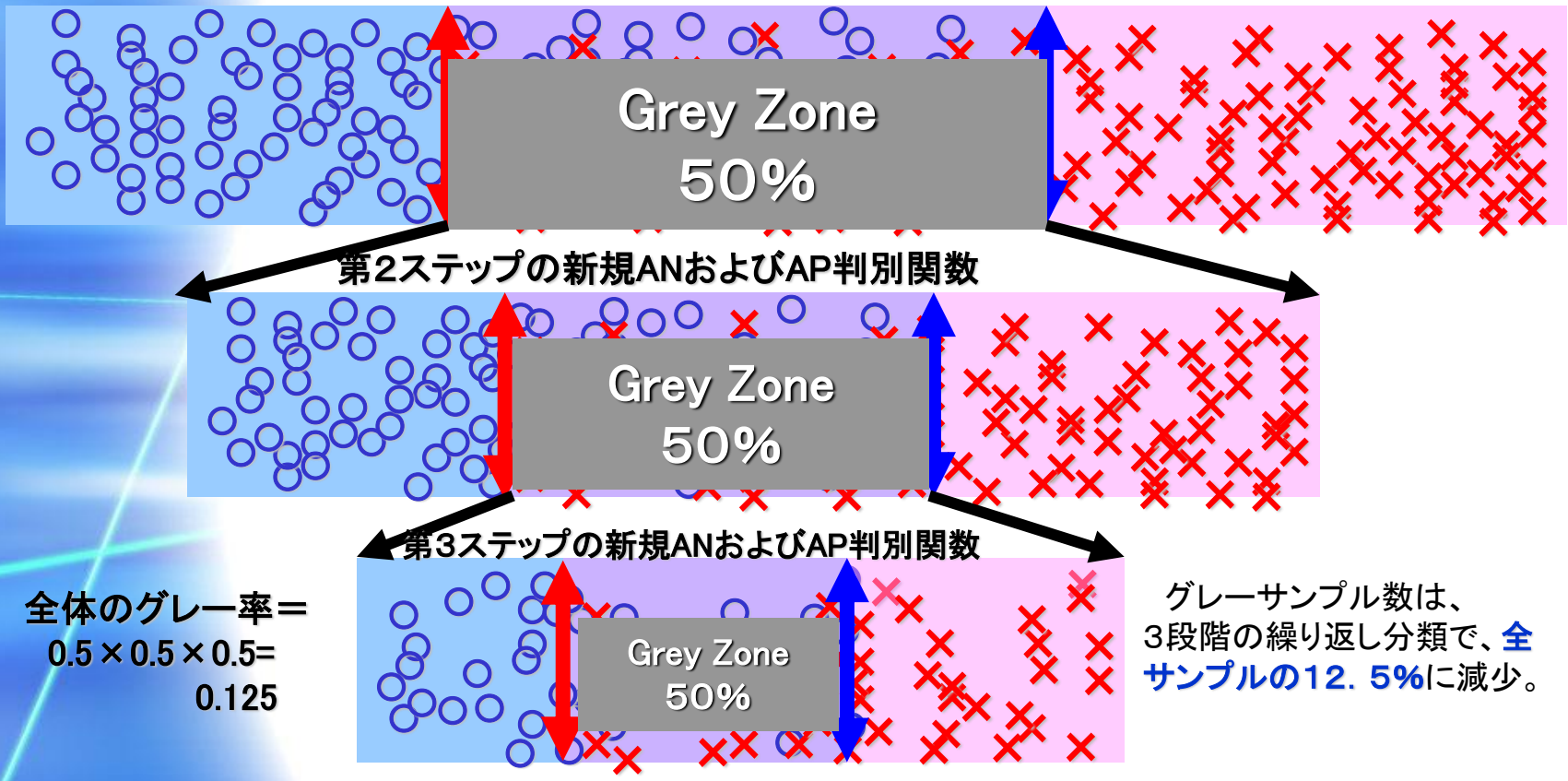
100%分類の実現は極めて困難

AP／AN判別関数の組み合わせによる分類特性



グレーゾーンを取り出して再分類を繰り返すアプローチ

“K-step Yard sampling（KY）法”



全体のグレー率＝
0.5×0.5×0.5=
0.125

グレーサンプル数は、
3段階の繰り返し分類で、**全サンプルの12．5%**に減少。

□ KY(K-step Yard sampling)法について

Challenge for classification and prediction

K-step Yard sampling methods

KY-methods

↓

The most powerful and advanced data analysis method

↓

The most difficult classification problem
6,965 sample of Ames test samples were,
Classified perfectly

Application test of "K-step Yard sampling"

☐Samples

1. Ames test data
2. Sample population
   total :6,965
   Mutagen; 2,932
   Non-mutagen; 4,033

☐Result of KY-method

1. Number of steps : 23 steps ; 22 (2 models) + 1 (1 model)
2. Classification ratio : 100 %

☐Used system

ADMEWORKS / ModelBuilder　V 3.0.22

☐Used parameters (Initial condition)

Number of generated parameters : 838
Number of parameters for step 1 : 98
Confidency index (Samples(6965) / Parameters(98)) : 71.1 > 4.0

Application test by normal and various D.A. methods

1. Linear discriminant analysis with linear least-squares method

Classification ratio : total; 73.50(6965), Mutagen;73.02(2932), Non mutagen;73.84(4033)
Number of mis-classified : (1846), (791) (1055)
Prediction ratio (L100 out) 72.58% deviance(0.92%)
(L500 out) 73.32% deviance(0.18%)

2. SVM (Support Vector Machine with Kernel)

Classification ratio : total; 90.87(6965), Mutagen;86.83(2932) Non mutagen; 93.80(4033)
Number of mis-classified : (636), (386) (250)
Prediction ratio (L500 out) 80.99% deviance(9.88%)

3. AdaBoost

Classification ratio : total; 77.24(6965), Mutagen;66.13(2932) Non-mutagen; 85.32(4033)
Number of mis-classified : (1585) (993) (592)
Prediction ratio (L500 out) 75.16% deviance(2.08%)

Classification result by the AdaBoost

Classification ratio of total 6,965 compounds is 77.24%

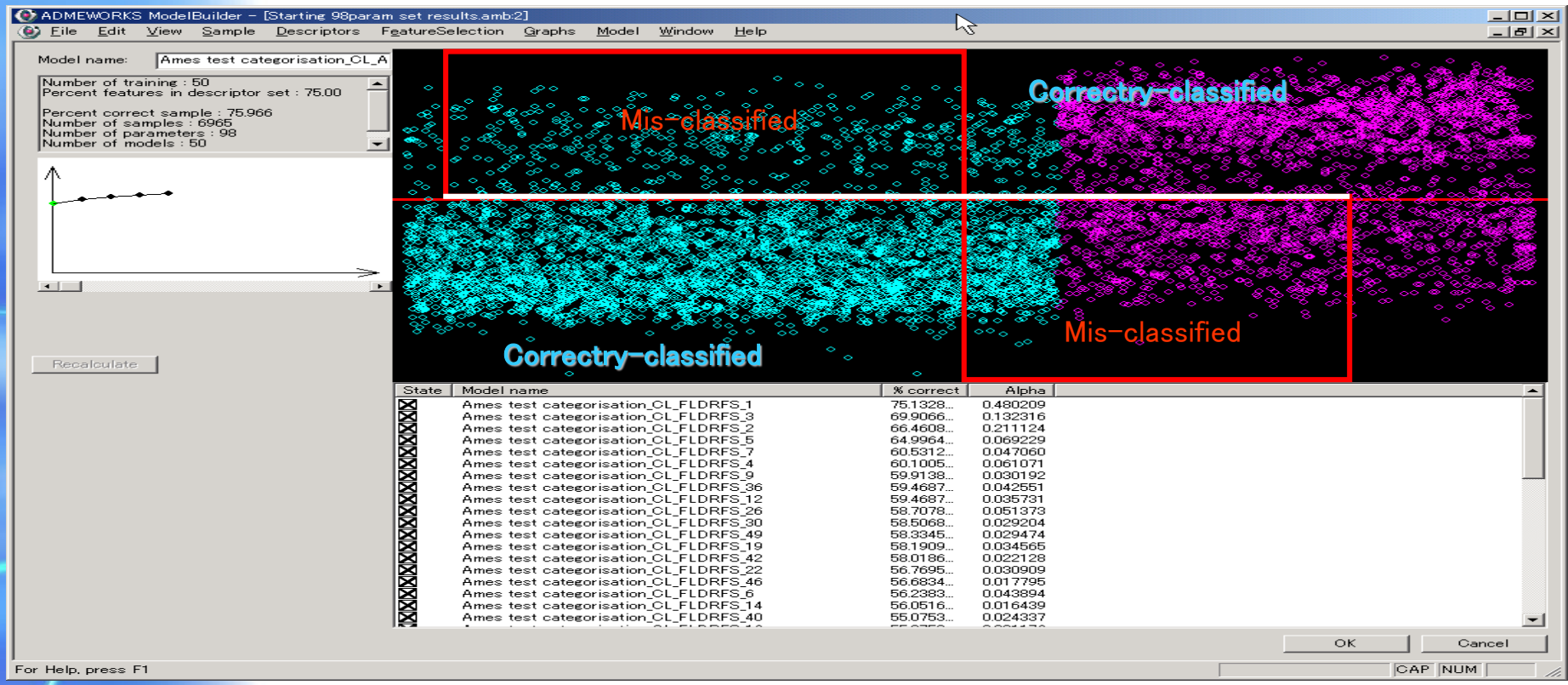
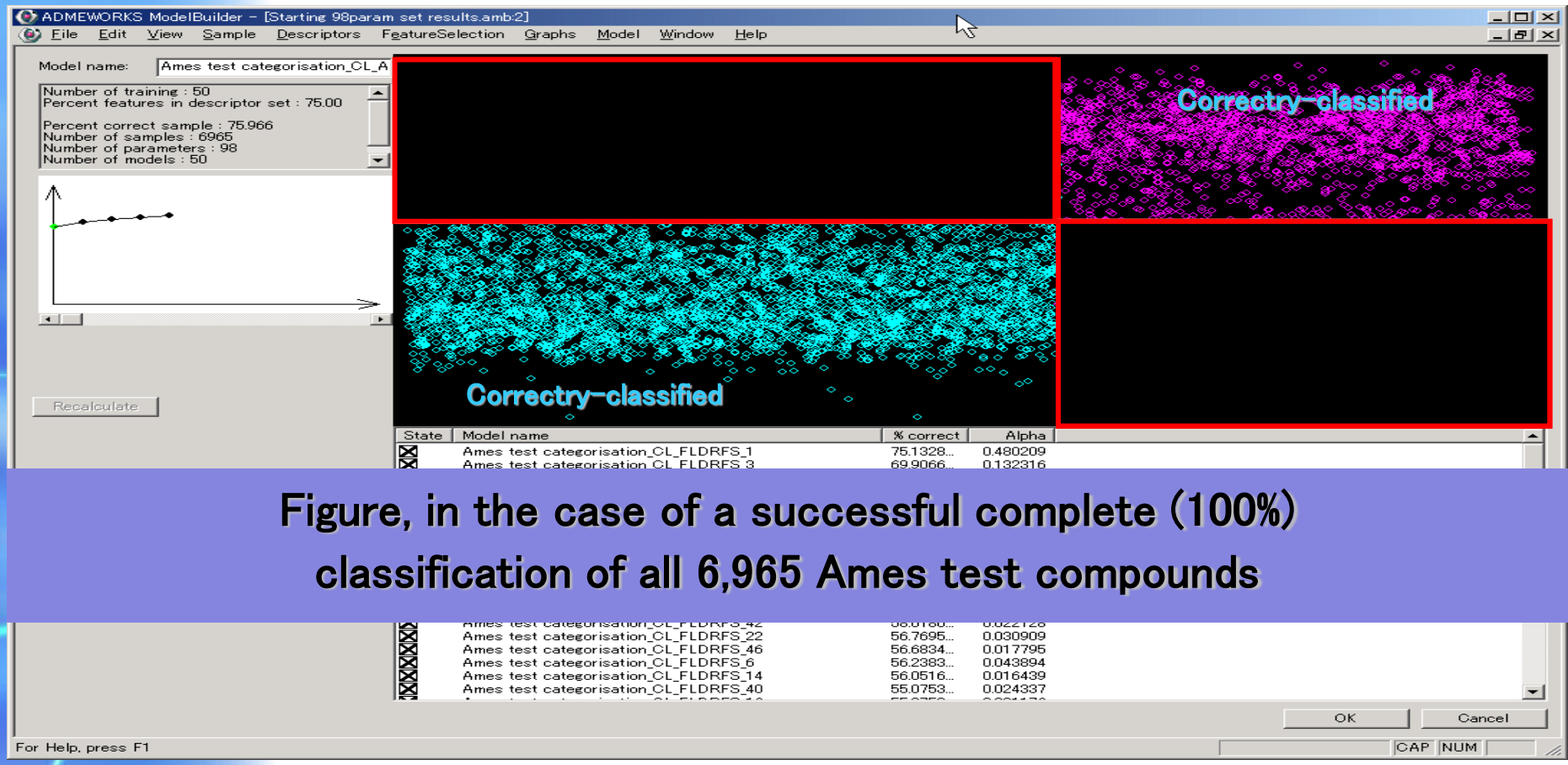


Figure of the ADMEWORKS

Classification result by the AdaBoost



Figure, in the case of a successful complete (100%) classification of all 6,965 Ames test compounds

Figure of the ADMEWORKS

“K-step Yard sampling (KY) Method”

Total steps : 23 steps (2 models) + 1 step (1 model)

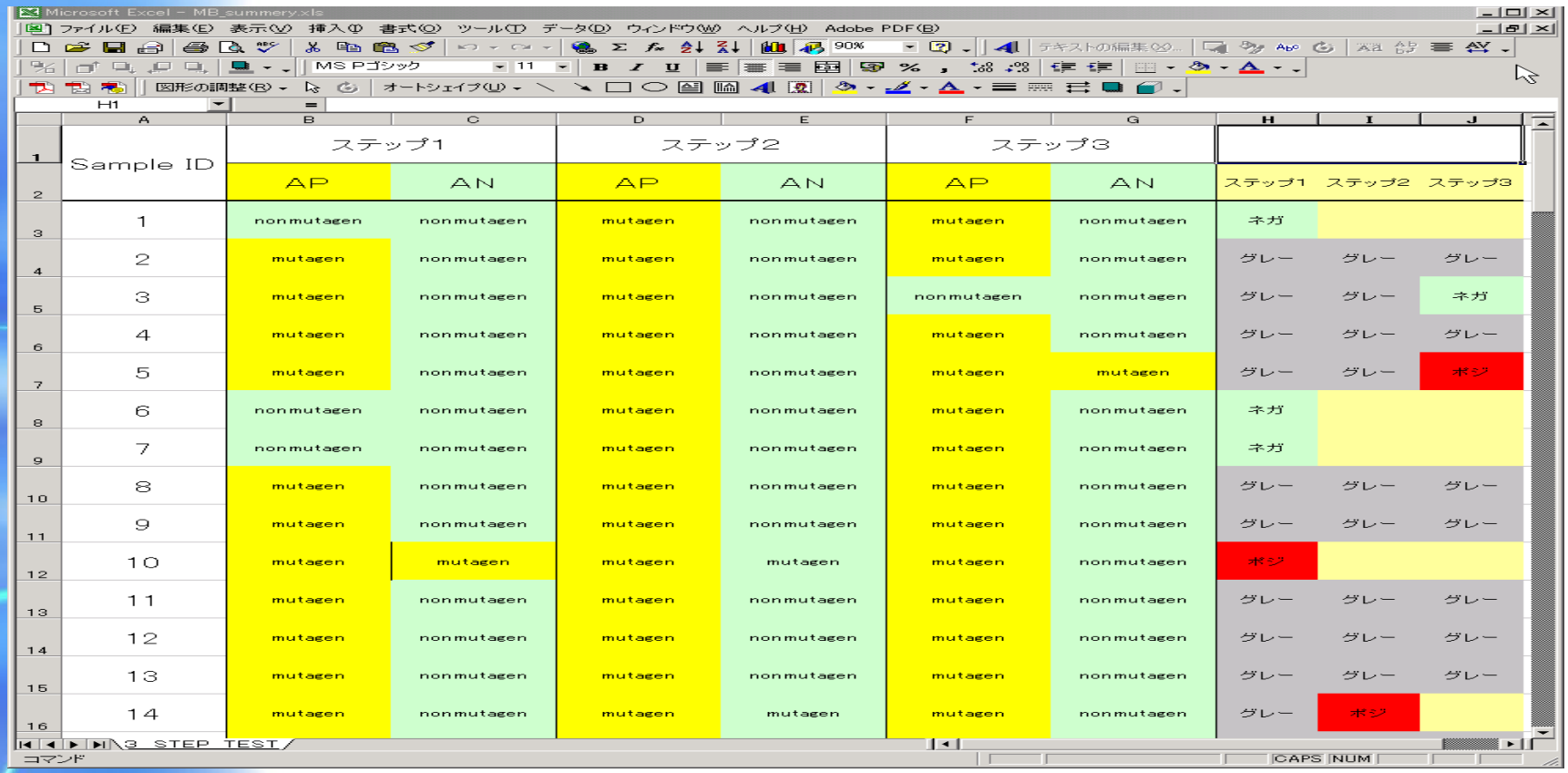
ステップID（KY法）	Starting samples(Total) / Final samples / Determined samples(Total)	Mutagen (Initial) / Mutagen (Final) / Determined samples(Mut.)	Non-mutagen (Initial) / Non-mutagen (Final) / Determined samples(Non-mut.)	Grey sample (Initial) / Grey sample (Final) / Grey ratio(%) (Grey/Total)
1	6965	2932	4033	0
	5864	2413	3451	5864
	1101	519	582	84.19
2	5864	2413	3451	5864
	5108	2142	2966	5108
	756	271	485	87.11
3	5108	2142	2966	5108
	4486	1919	2567	4486
	622	223	399	87.82
4	4486	1919	2567	4486
	4133	1779	2354	4133
	353	140	213	92.13
5	4133	1779	2354	4133
	3794	1651	2143	3794
	339	128	211	91.8
6	3794	1651	2143	3794
	3462	1485	1977	3462
	332	166	166	91.25
7	3462	1485	1977	3462
	3090	1345	1745	3090
	372	140	232	89.25
8	3090	1345	1745	3090
	2826	1220	1606	2826
	264	125	139	91.46
9	2826	1220	1606	2826
	2592	1139	1453	2592
	234	81	153	90.63
10	2592	1139	1453	2592
	2384	1047	1337	2384
	208	92	116	91.98

“K-step Yard sampling (KY) Method”

12	2095	931	1164	2095
	1848	829	1019	1848
	247	102	145	88.21
13	1848	829	1019	1848
	1607	733	874	1607
	241	96	145	86.96
14	1607	733	874	1607
	1380	623	757	1380
	227	110	117	85.87
15	1380	623	757	1380
	1028	466	562	1028
	352	157	195	74.49
16	1028	466	562	1028
	787	358	429	787
	241	108	133	76.56
17	787	358	429	787
	529	234	295	529
	258	124	134	67.22
18	529	234	295	529
	392	201	191	392
	137	33	104	74.1
19	392	201	191	392
	279	141	138	279
	113	60	53	71.17
20	279	141	138	279
	184	105	79	184
	95	36	59	65.95
21	184	105	79	184
	112	66	46	112
	72	39	33	60.87
22	112	66	46	112
	66	39	27	66
	46	27	19	58.93
	66	39	27	66
23(1 model)	0	0	0	0
	66	39	27	0

“K-step Yard sampling (KY) Method”

Classification results by 3 steps



Sample ID	ステップ1		ステップ2		ステップ3				
	AP	AN	AP	AN	AP	AN	ステップ1	ステップ2	ステップ3
1	nonmutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	ネガ		
2	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
3	mutagen	nonmutagen	mutagen	nonmutagen	nonmutagen	nonmutagen	グレー	グレー	ネガ
4	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
5	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	mutagen	グレー	グレー	ポジ
6	nonmutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	ネガ		
7	nonmutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	ネガ		
8	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
9	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
10	mutagen	mutagen	mutagen	mutagen	mutagen	nonmutagen	ポジ		
11	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
12	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
13	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
14	mutagen	nonmutagen	mutagen	mutagen	mutagen	nonmutagen	グレー	ポジ	

Spatial features of "K-step Yard sampling"

☐Summary 1

■Features

1. KY-method is a meta-method
   Executed on various DA (linear and non-linear) methods.

2. Perfect classification is achieved on any condition

3. Three different types of KY-methods are available at this time

4. Applicable not only on Binary classification but Fitting methods

Spatial features of "K-step Yard sampling"

☐Summary 2

■Advantages

1. Sample number free approach

2. Sample distribution free approach

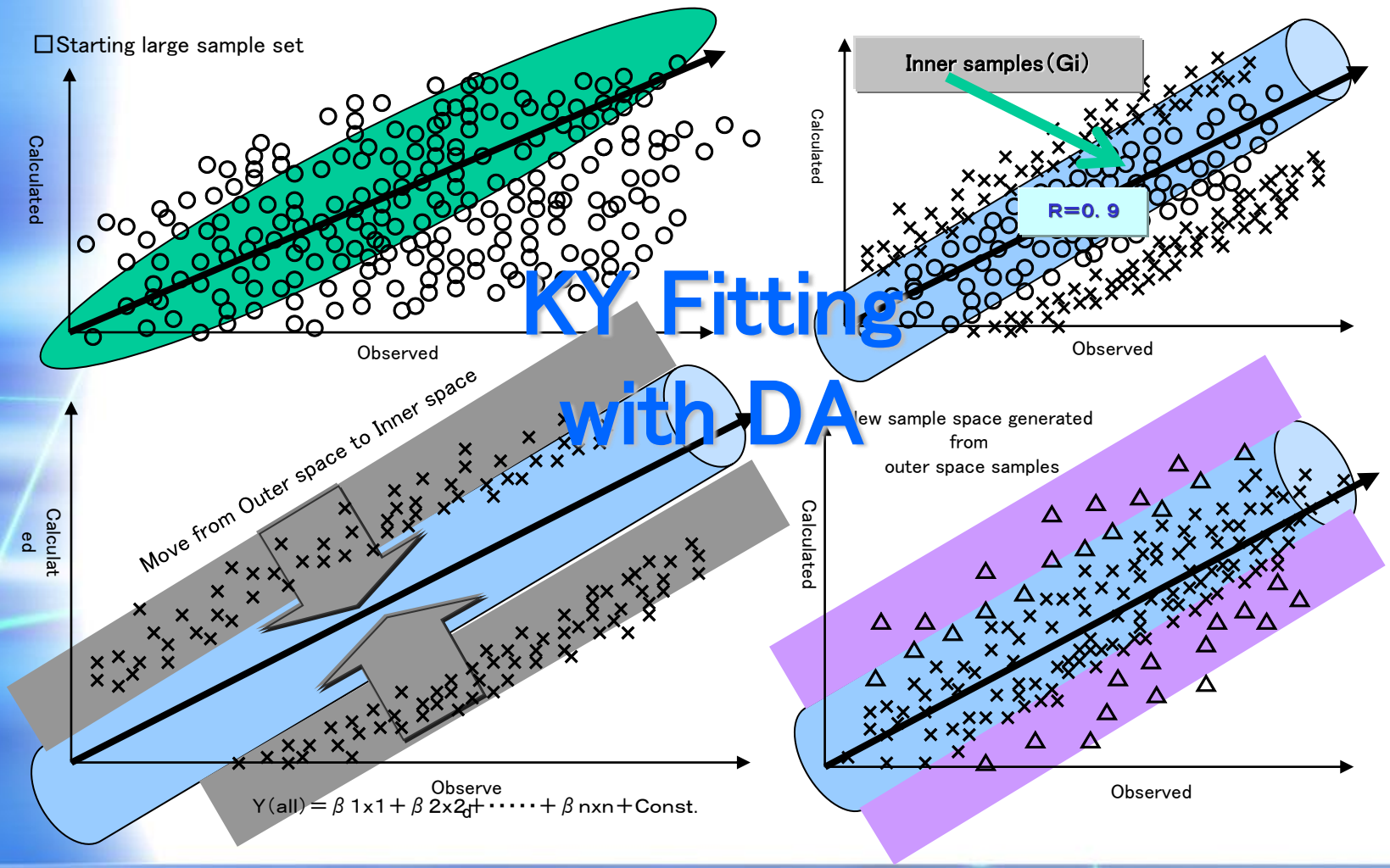
3. Perfect classification is achieved in any condition

4. Need not spatial data analysis softwares

■Disadvantages

Relatively complex operation to generate discriminant functions

◆KY method for fitting methods



◆KY method for fitting methods

Fish: 96 hours LC50、 Number of samples: 791、Log(1/LC50_Mm) (Max/Min) : 6.376 / −2.963

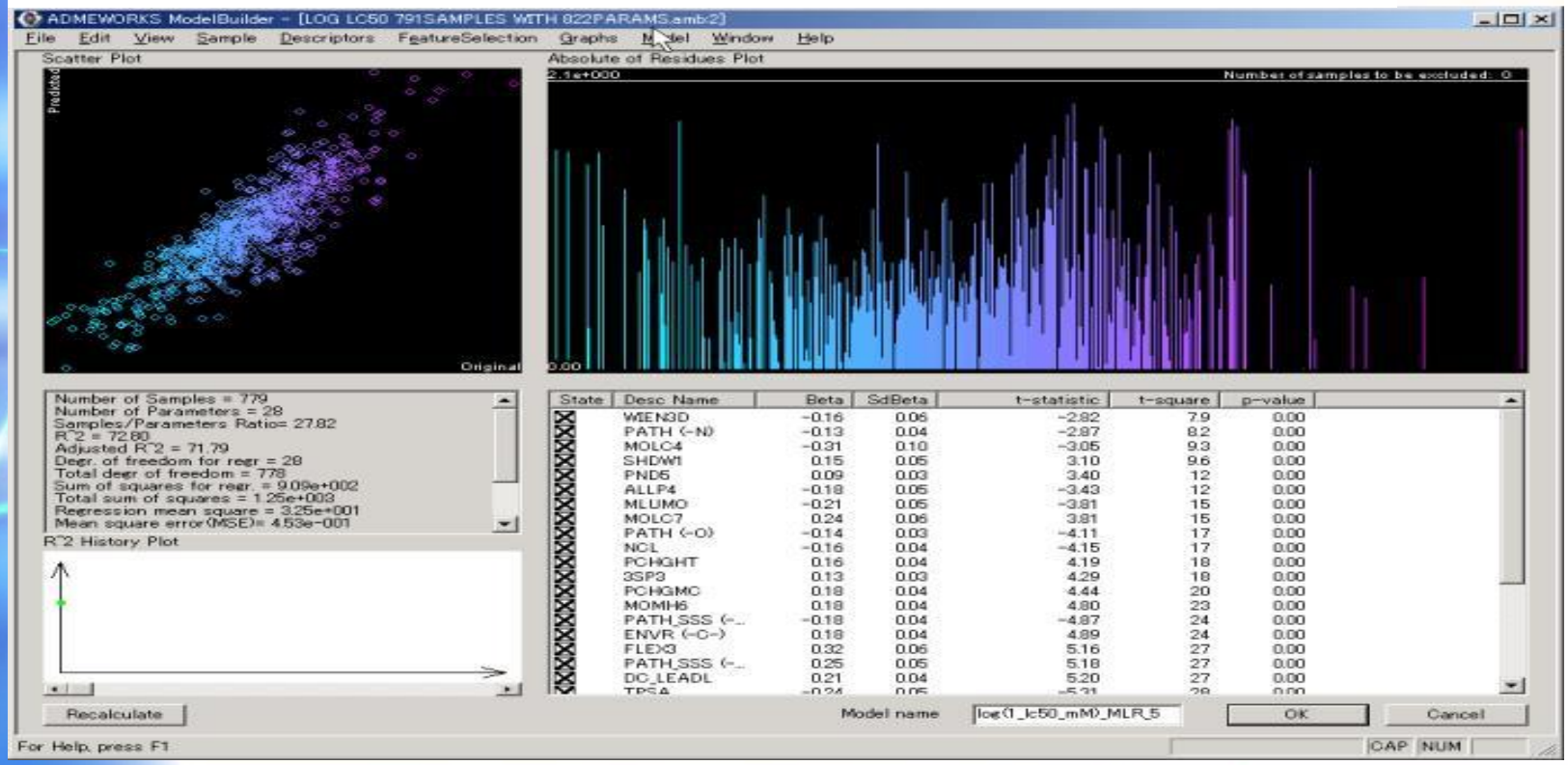
◇ Data analysis by ordinal linear

Step1 : Inner sample set

Number of samples : 779, Number of used parameters : 28, Confidance ratio : 27.8

R2 : 72.8, R : 85.3, F−value : 71.7, CV : 69.6

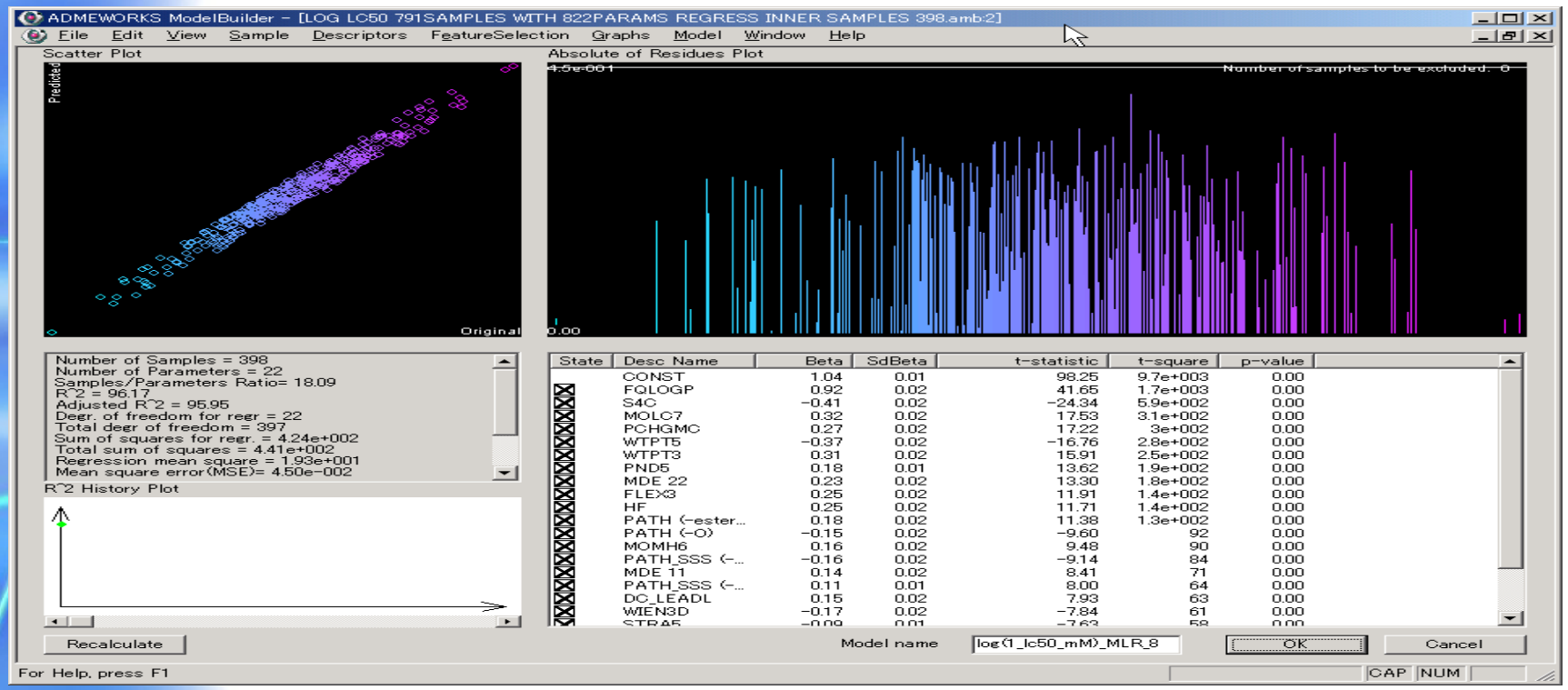
69. 6



◆フィッティングKY法実証実験（2）

◇フィッティングKY法による解析（ステージ1）

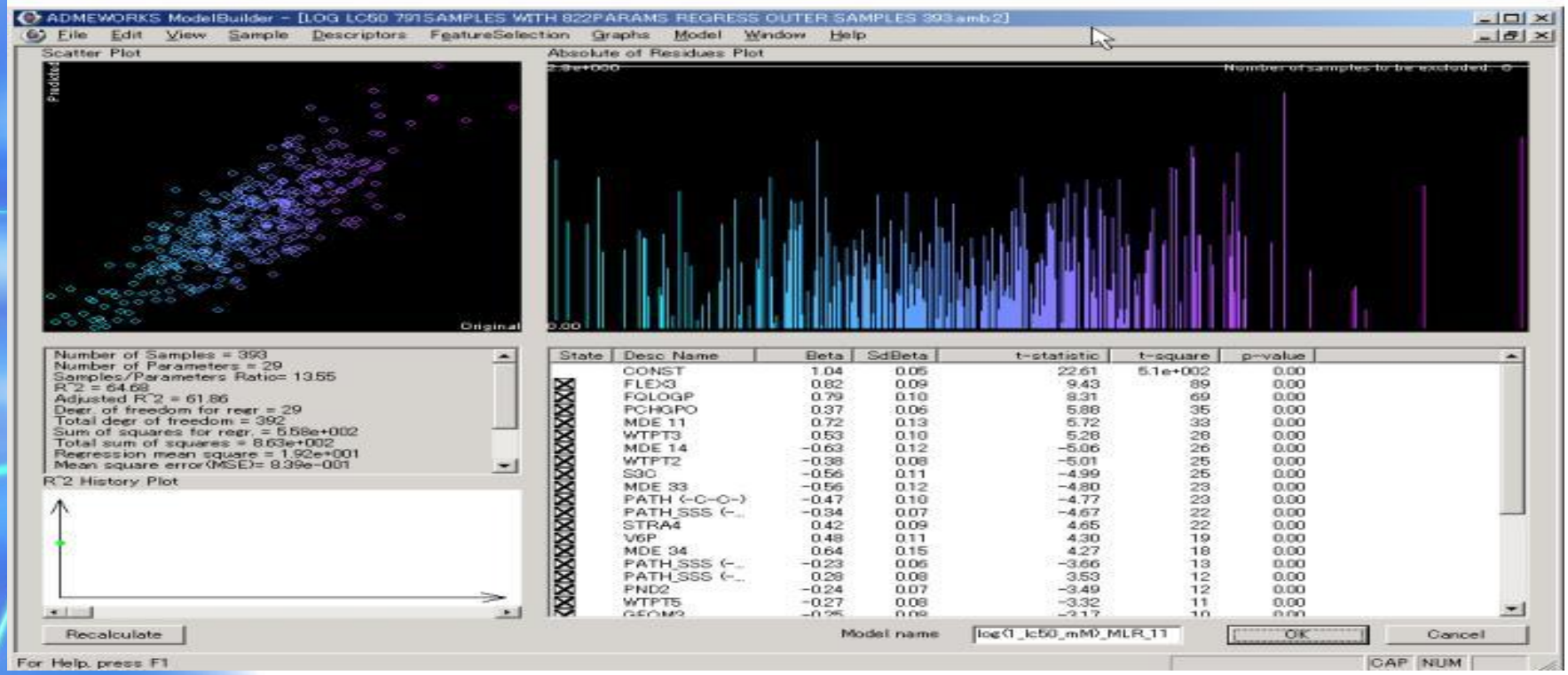
ステップ1：インナーサンプル
サンプル数：398、 パラメータ数：22、 信頼性指標：18．1
R2：96．2、 R：98．1、 F値：428、 クロスバリデーション：94．4



◆フィッティングKY法実証実験（3）

◇フィッティングKY法による解析（ステージ1）

ステージ1：　アウターサンプル
サンプル数：393、　パラメータ数：29、　信頼性指標：13．6
R2：64．7、　R：80．4、　F値：22.9、　クロスバリデーション：57．5



◇テーラーメードモデリング

＊予測モデルの精度は、モデル作成に用いたサンプルの質に大きく依存
＊予測モデルの精度向上にはサンプル母集団内のサンプルを目的解析に必要な情報を多く含むことが理想
＊解析目的と関係のある情報を含まないサンプルは極力避けるのが良い

＊化合物分野には「**似た化合物は似た活性／物性／毒性を有する**」という概念
＊予測対象化合物と似た化合物を選択して予測モデルを構築する

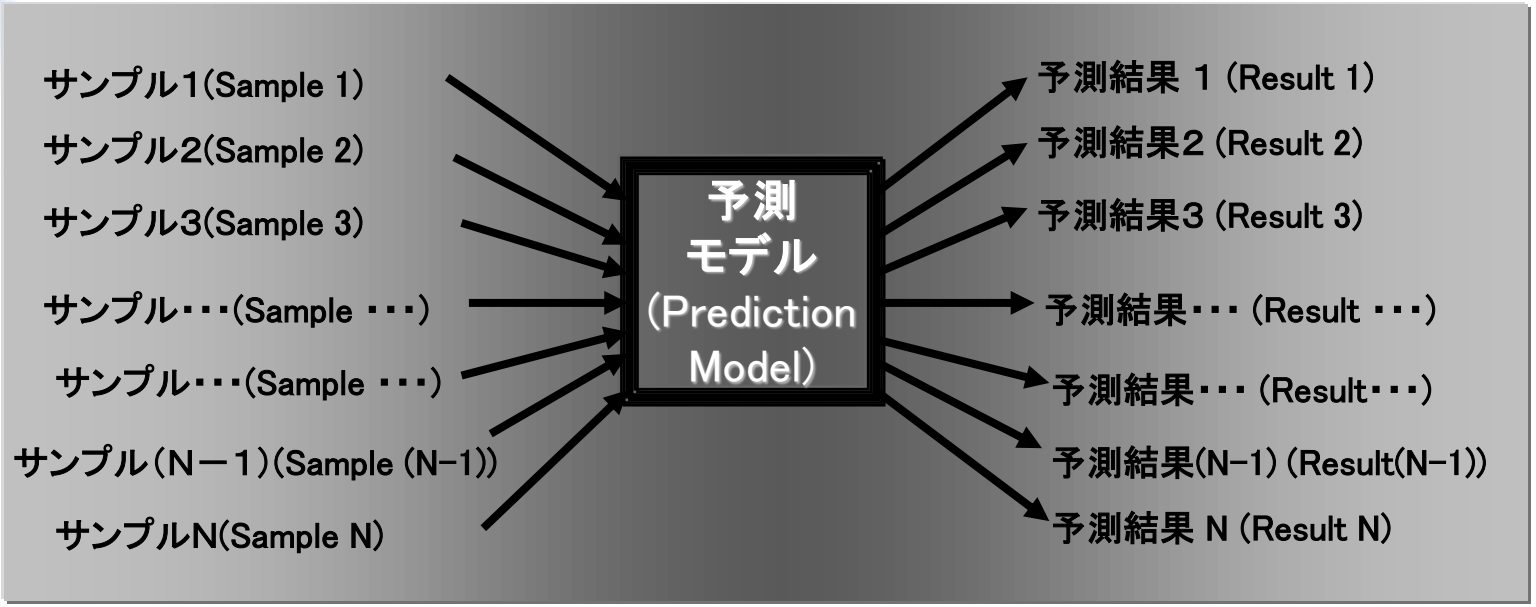
＊予測対象サンプル毎に、サンプル母集団を構成し、予測モデルを構築する

テーラーメードモデリング

United States Patent Application 20100145896 ; *Yuta June 10, 2010

従来手法による予測アプローチ
(Prediction approach by traditional method)

特徴：総てのサンプルを対象とした予測モデルの構築
Features:Generate a prediction model which can handle all samples



利点 (Merit):少ない数の予測モデル作成（Small number of prediction models are generated ）

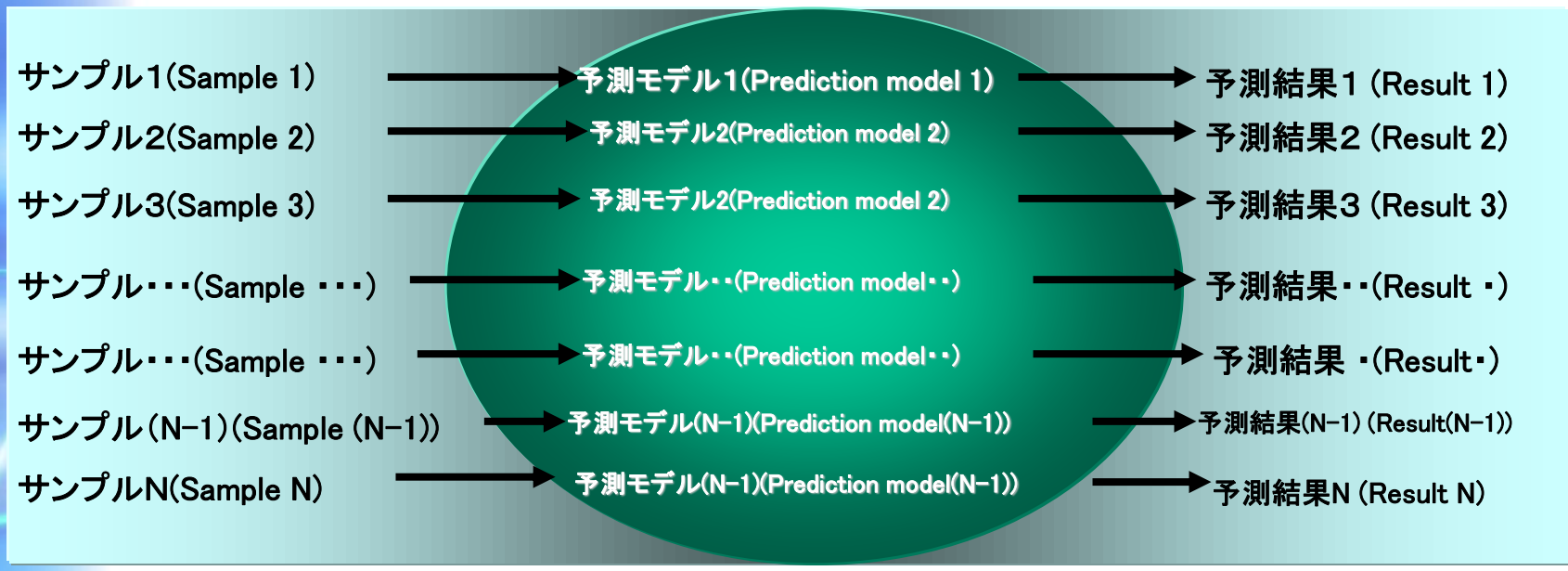
難点（Weakness）： 予測率の向上が困難である（Difficult to achieve high prediction ratio ）

「テーラーメード・モデリング」による予測アプローチ

(Prediction approach by “Tailor-Made Modeling”)

特徴:サンプル単位での予測モデルの構築

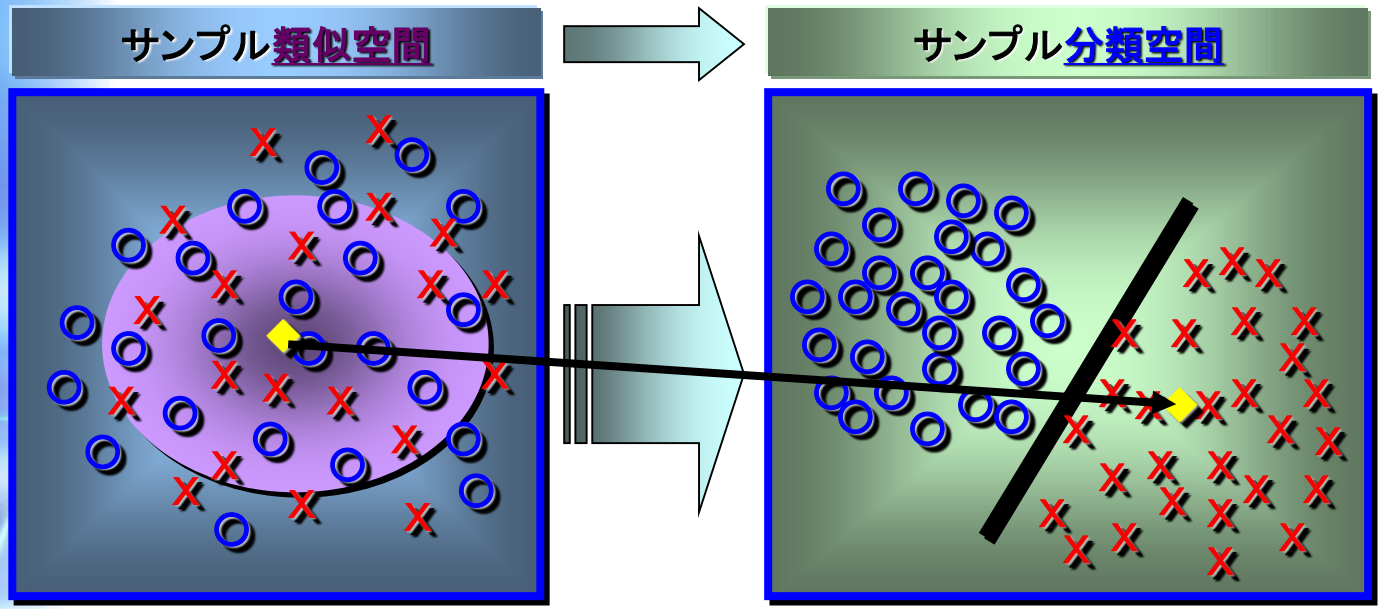
Features:Generate a prediction model which is designed for only 1 samples



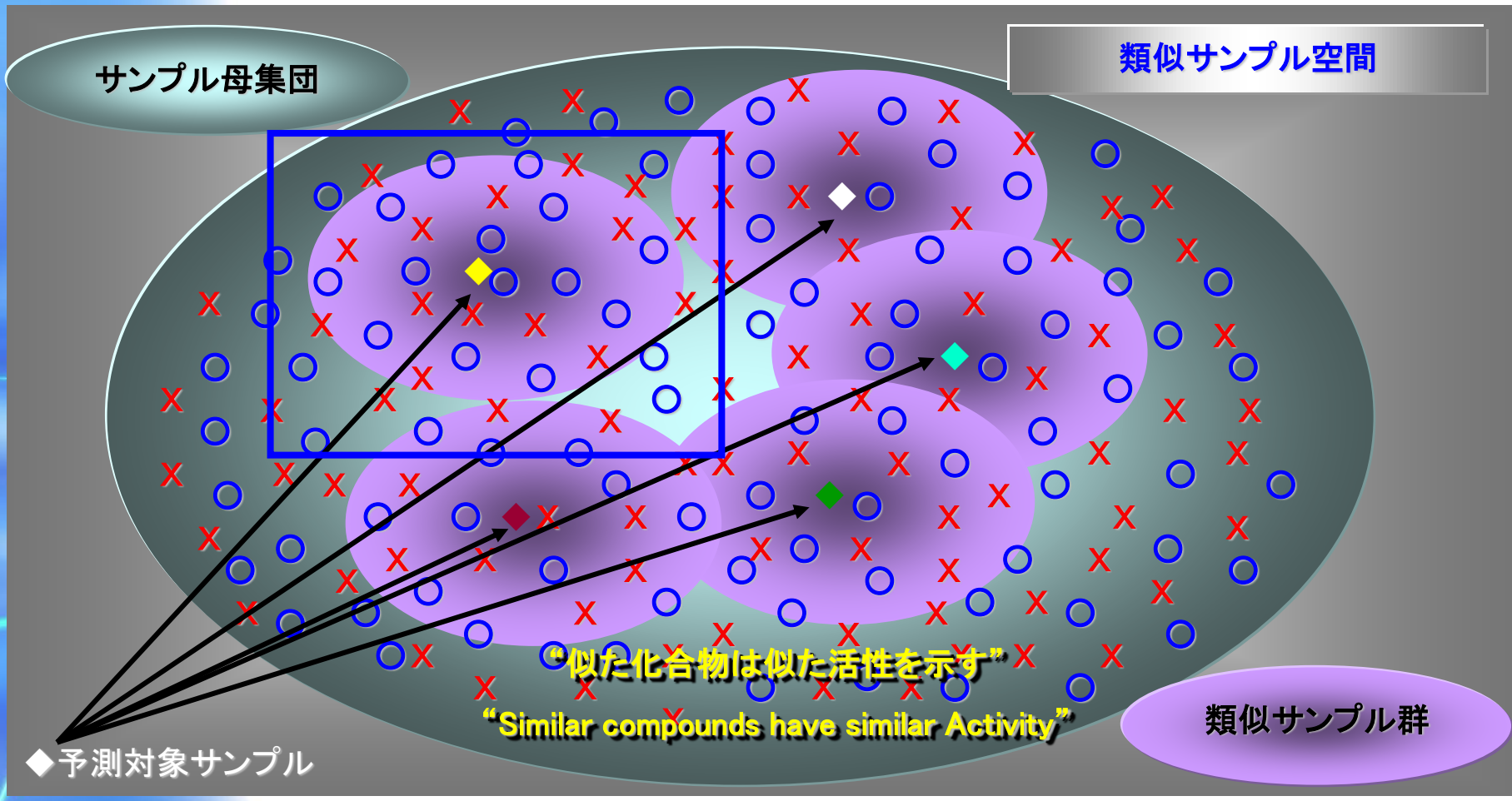
利点 (Merit)　　:　予測率が大幅に向上する (High prediction ratio will be achieved)

難点 (Weakness):　計算時間がかかる (Need large calculation time)

予測用サンプルの取り出しと、テーラーメード予測



サンプル母集団からの予測用サンプルの取り出し



◇ KY(K-step Yard sampling)法

■ **ニクラス分類KY法:Binary classification KY methods**
- ニモデルKY法; Two models BC KY method
- 一モデルKY法; One model BC KY method
- モデルフリーKY法; Model free BC KY method

■ **重回帰(フィッティング)KY法:Regression KY methods**
- 判別関数付き重回帰KY法;BC regression KY method
- 三ゾーン重回帰KY法;Three zone regression KY method
- モデルフリー重回帰KY法;Model free regression KY method

□ クラスタリングKY法; Clustering KY methods
□ 主成分KY法; Principal component KY methods

◇ 現在開発中
- リバーシKY法:ニクラス分類の簡略法
- ポピュレーションフリーKY法:ポピュレーション比率の悪い時に適用

◇ テーラーメードモデリング(Tailor made modeling)

■は特許出願済み、□は検証済み。なお、テーラーメードモデリング特許出願済