

CBI学会2018年大会
フォーカストセッション FS-03

「ケモメトリックス」の基礎

株式会社 インシリコデータ
湯田 浩太郎

<http://www.insilicodata.com>

ケモメトリックスとデータサイエンス ビッグデータ、人工知能等

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

■ケモインフォマティクス(cheminformatics):

コンピュータと情報化技術を用いて、化学領域の問題に適用する方法論である

■ケモデータサイエンス、化学データサイエンス by Yuta
(chemodata science, chemical data science):

化学分野の諸問題についてデータサイエンス手法を適用して解決を目指す学問分野である

関連用語:

ビッグデータ(big data)、
人工知能(artificial intelligence)、
プログラミング言語

ケモメトリックスの適用分野

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

並列創薬

インシリコンビ

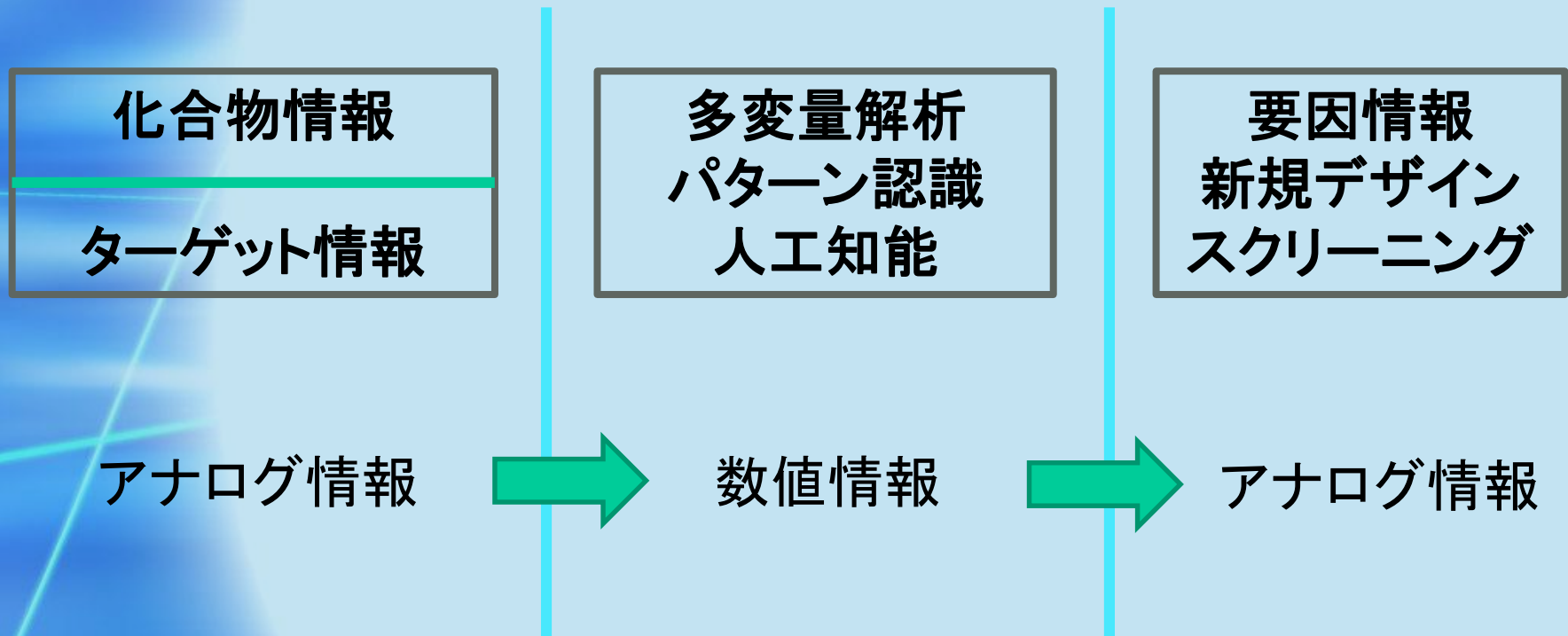
医療解析

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



自動車関連事項を基本としたケモメトリックス技術構成内容比較

自動車本体

設計、組み立て、
機構メカニズム、
車種、デザイン

データ解析手法論
クラス分類・予測
クラスタリング
マッピング
その他

移動結果・目的

目的地
道路地図
行動目的の実施

QSTR、QSAR、
QSPR、要因解析、
新規化合物デザイン、
バーチャルスクリーニ
ング、ドラグリポジショ
ニング、他

ケモメトリックスの議論点

自動車関連事項を基本としたケモメトリックス技術構成内容比較

自動車本体

設計、組み立て、
機構メカニズム、
車種、デザイン

自動車操作関連

道路交通法
運転技術
走行条件
自動車保険

移動結果・目的

目的地
道路地図
行動目的の実施

データ解析手法論
クラス分類・予測
クラスタリング
マッピング
その他

QSTR、QSAR、
QSPR、要因解析、
新規化合物デザイン、
バーチャルスクリーニ
ング、ドラグリポジショ
ニング、他

自動車関連事項を基本としたケモメトリックス技術構成内容比較

自動車本体

設計、組み立て、
機構メカニズム、
車種、デザイン

自動車操作関連

道路交通法
運転技術
走行条件
自動車保険

移動結果・目的

目的地
道路地図
行動目的の実施

データ解析手法論
クラス分類・予測
クラスタリング
マッピング
その他

手法の適用留意事項
化合物関連操作
サンプル環境
パラメーター環境

QSTR、QSAR、
QSPR、要因解析、
新規化合物デザイン、
バーチャルスクリーニ
ング、ドラグリポジショ
ニング、他

■解析原理・化合物操作・データ解析での留意事項

多変量解析／パターン認識によるデータ解析⇒宝箱・発見型アプローチ

化合物関連：一元一項対応、一次元表記、二次元表記、三次元表記

三次元の扱い問題：ローカル・グローバルミニマ対応

確率問題⇒偶然相関

フィッティング⇒オーバーフィッティング（過剰適合）

線形/非線形問題；

特徴抽出（パラメーター手法）；種類、特徴、限界

サンプル関連；総サンプル数、最小サンプル数、

ポピュレーション（絶対数、クラス比率）

パラメーター関連：種類、単位の違い、オートスケーリング、最少パラメーター数

ニューラルネットワーク⇒中間層の数 ⇒ 縦の重なり

・問題点；チャンスコリレーション、非線形性

パラメーター数が多いとき：パラメーター圧縮 PCA PLS

パラメーター数が少ない時；成功率低下、解析不能、物性式等

要因解析：パラメーターの読解力、分類力が強いパラメーター

KY法：サンプル数フリー、ポピュレーションフリー

二クラス分類；完全分類

重回帰；高相関係数、高絶対係数

分類/予測⇒クロスバリデーション

□ ケモトリックス適用の基本原則

種々データ解析でのケモ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

- * 仮説検証型アプローチ(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

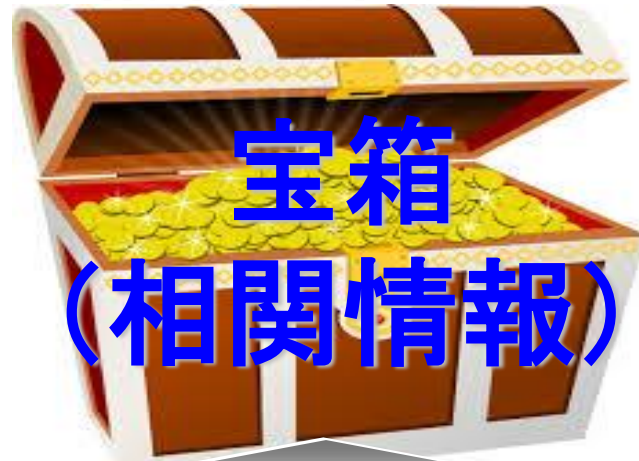
様々な重要情報を仮説抜きで発見(探索)する

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

多変量解析やパターン認識適用の基本原則

情報等価原理

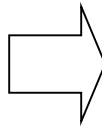
化合物
蛋白
ゲノム



薬理活性
毒性
副作用
代謝
分解性

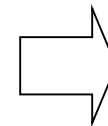
情報等価性／必然性

原因情報



説明義務

パターン認識手法

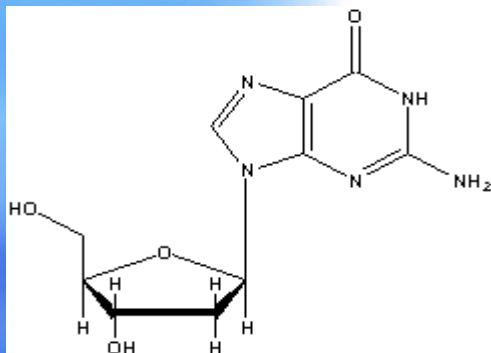


結果情報

多変量解析やパターン認識適用の流れ(1)

(化合物構造式と目的活性データを用いた予測)

化合物構造式



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

情報の等価性／必然性

最終
パラメータ群

宝箱
(相関情報)

薬理活性
毒性
副作用
代謝
分解性

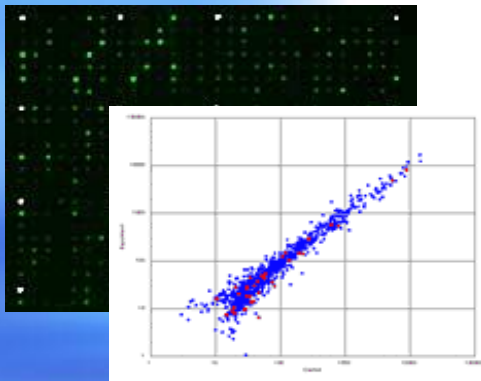
種々の解析手法

- ・判別分析
- ・クラスタリング
- ・フィッティング
- ・その他

多変量解析やパターン認識適用の流れ(2)

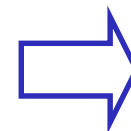
(発現プロフィールデータを用いた予測)

発現プロフィールデータ



数値パラメータ

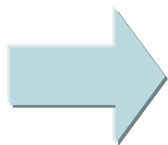
Hs.35092 ;5523
Hs.27935 ; 857
Hs.622 ;3682
Hs.38677 ;2283
.....



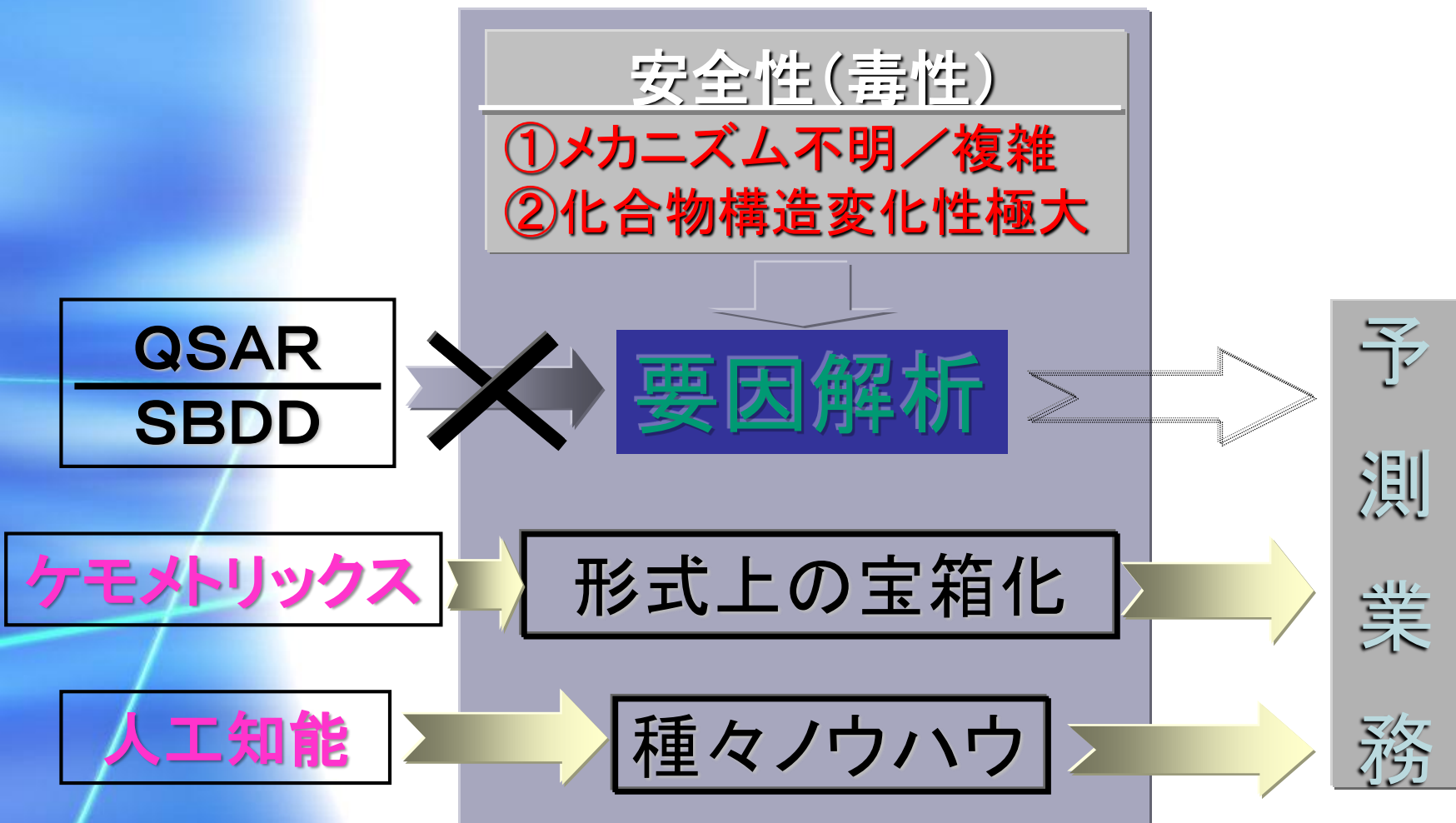
薬理活性
毒性
副作用
代謝
分解性

種々の解析手法

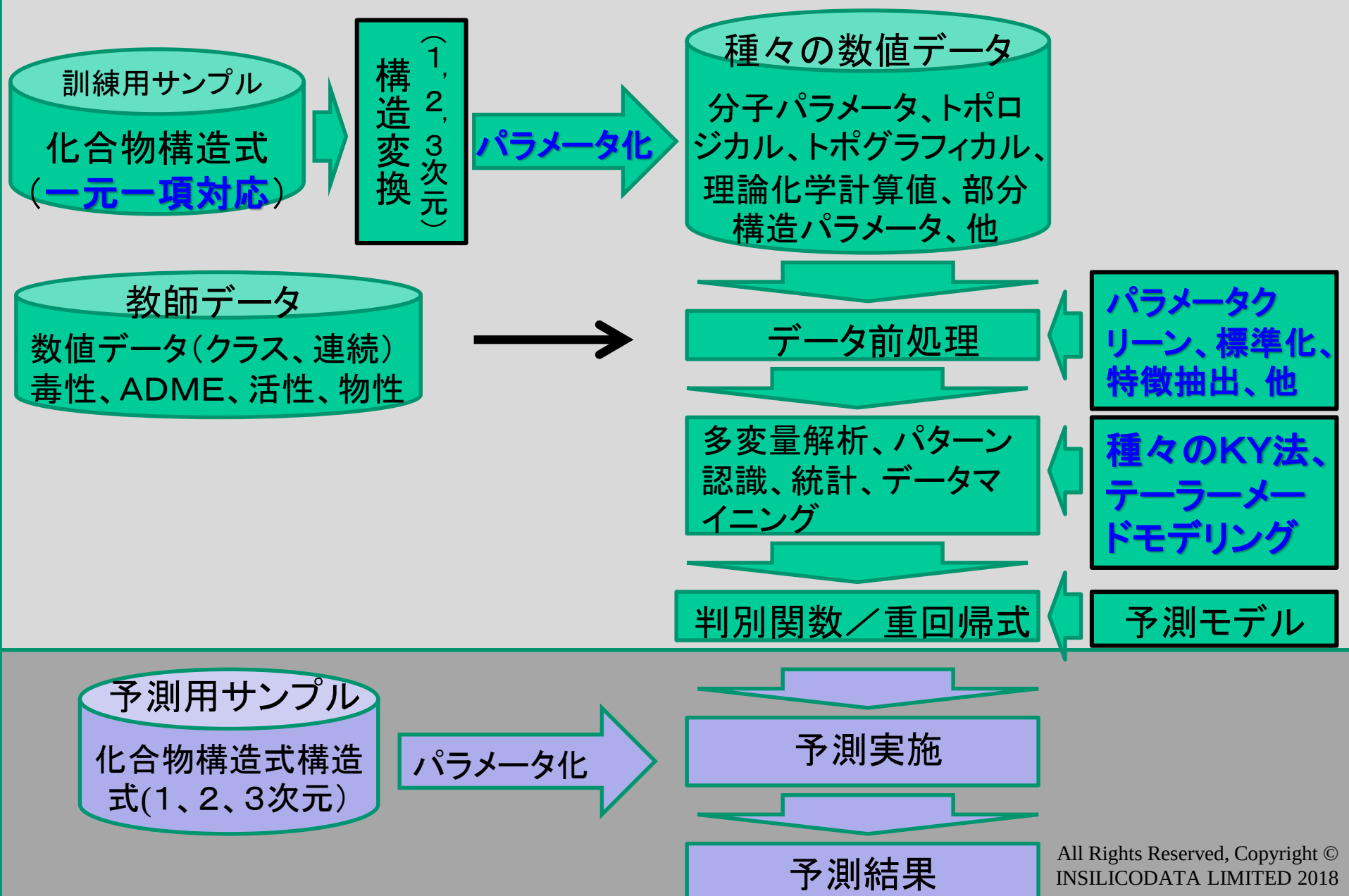
判別分析
クラスタリング
マッピング
フィッティング
その他



コンピュータによる安全性予測へのアプローチ

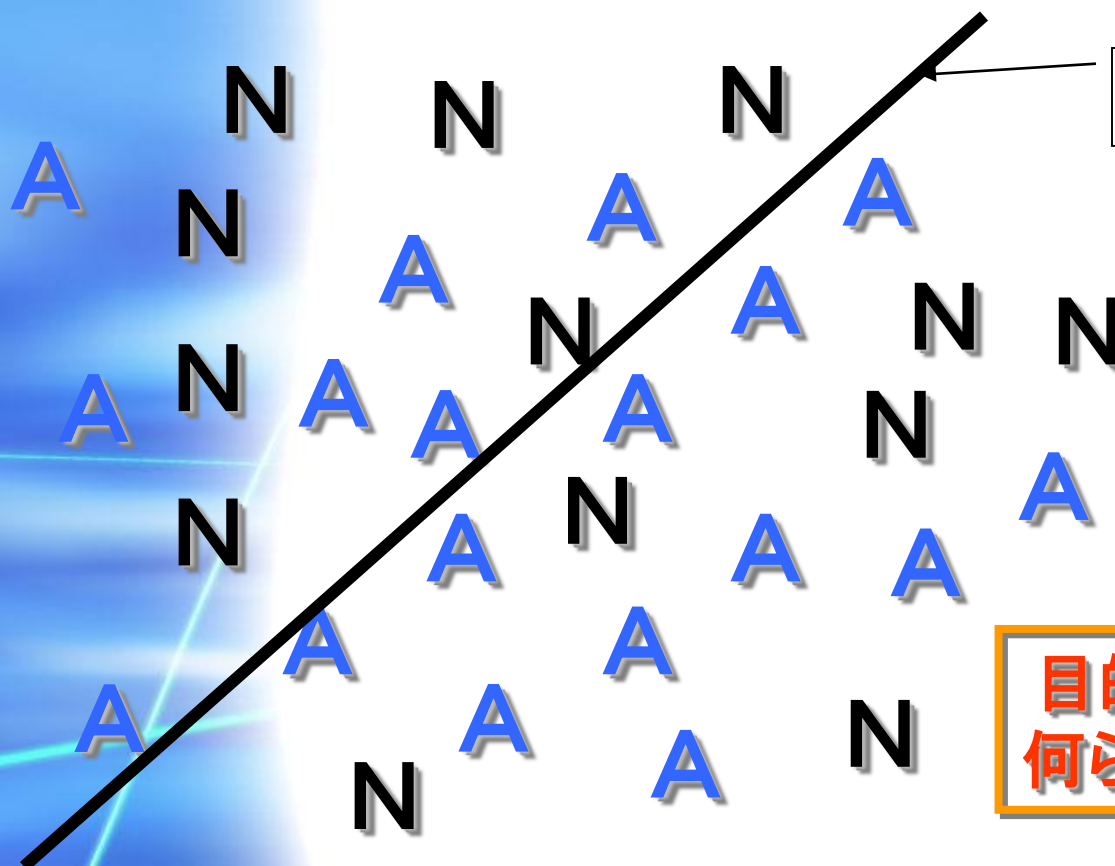


ケモメトリックス研究支援システム構成図



□ 判別関数による二クラス分類の基本

■ ノイズデータを含んだN次元パターン空間



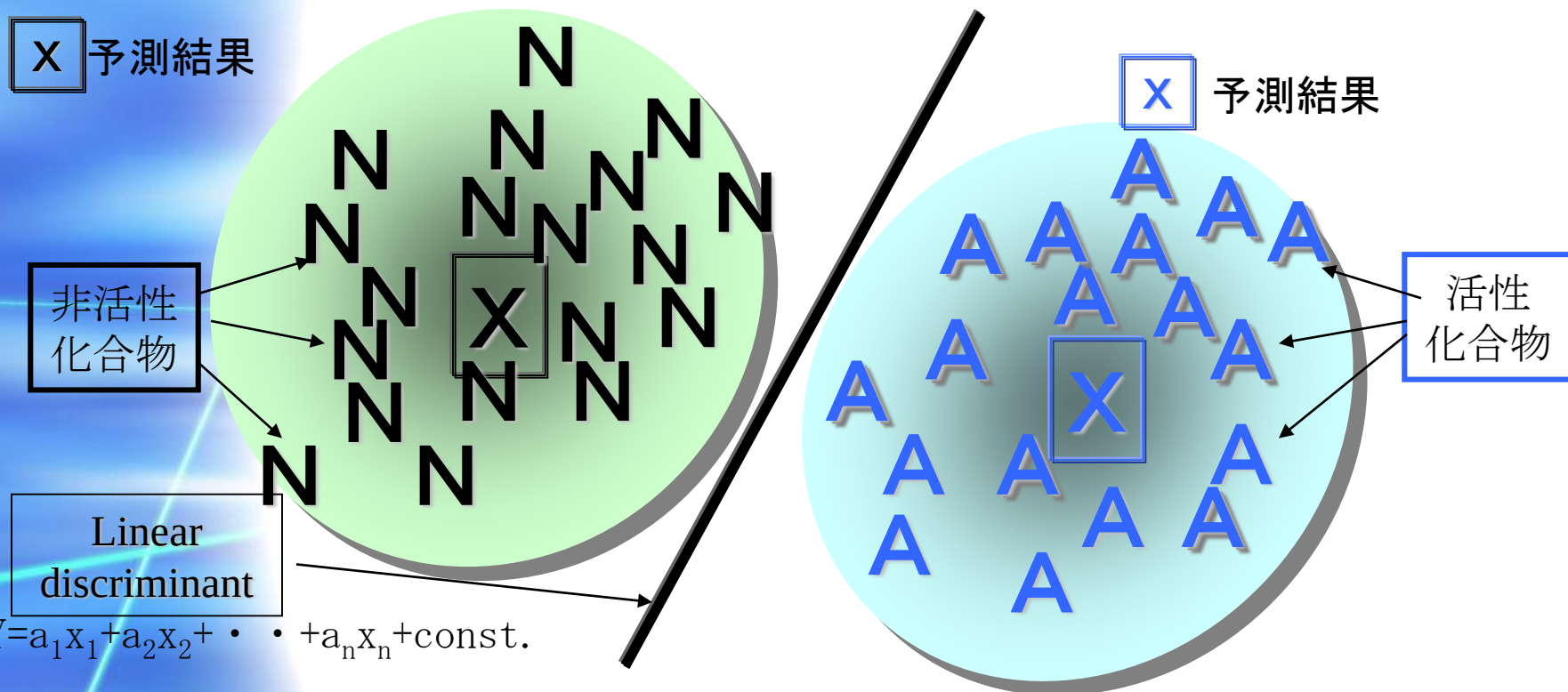
線形判別関数

$$Y = a_1 x_1 + a_2 x_2 + \dots + a_n x_n + \text{const.}$$

目的活性とパターン空間に
何ら相関関係が見出せない

A ; 活性化合物
N ; 非活性化合物

■ 目的活性と相関の高いパラメータ群により 形成されるN次元空間



* 活性および非活性化合物群と2分割可能なパターン空間

* 本空間を作るパラメータ群には、活性／非活性と分類するに重要な情報（科学的根拠／相関）を持つ

多変量解析／パターン認識による 相関解析で得られる情報

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

Y: 目的活性／ADME-T／他

$$Y \geq 0$$

- ・代謝活性あり
- ・毒性あり

$$Y < 0$$

- ・代謝活性無し
- ・毒性無し

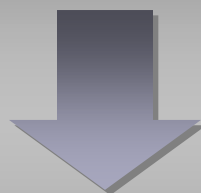
相関要因の種類

相関の強さ

相関の方向性

□構造－代謝／毒性／他との要因解析

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

構造－薬理活性／ADME-T／物性相関

データ解析特性の理解

□ 線形解析と非線形解析⇒過剰適合の危険性

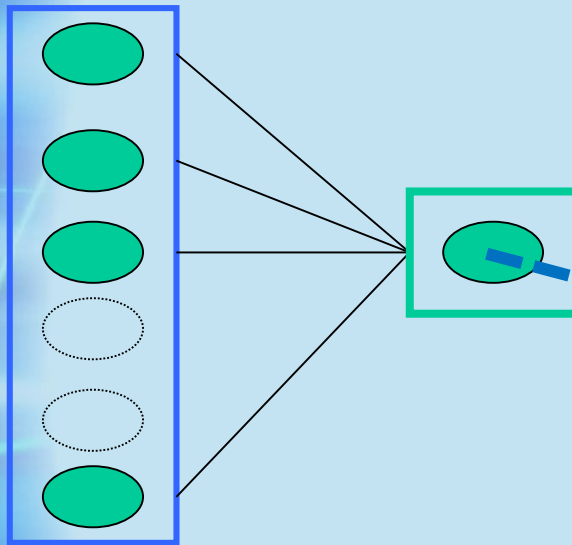
- ・データ解析上での重要性
- ・データ解析実施上での留意点
- ・解析手法のイメージから線形と非線形解析を理解する
- ・線形および非線形解析の予測に関する違い

ニューラルネットワーク:パーセプトロン

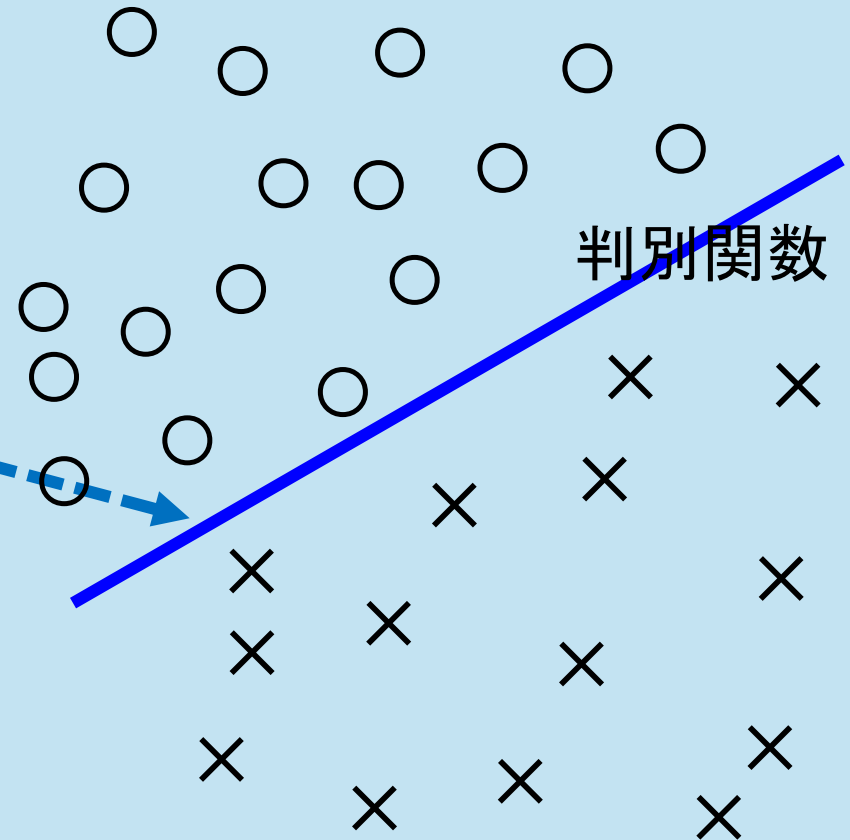
ネットワーク構造による

線形クラス分類のイメージ図

入力パラメーター

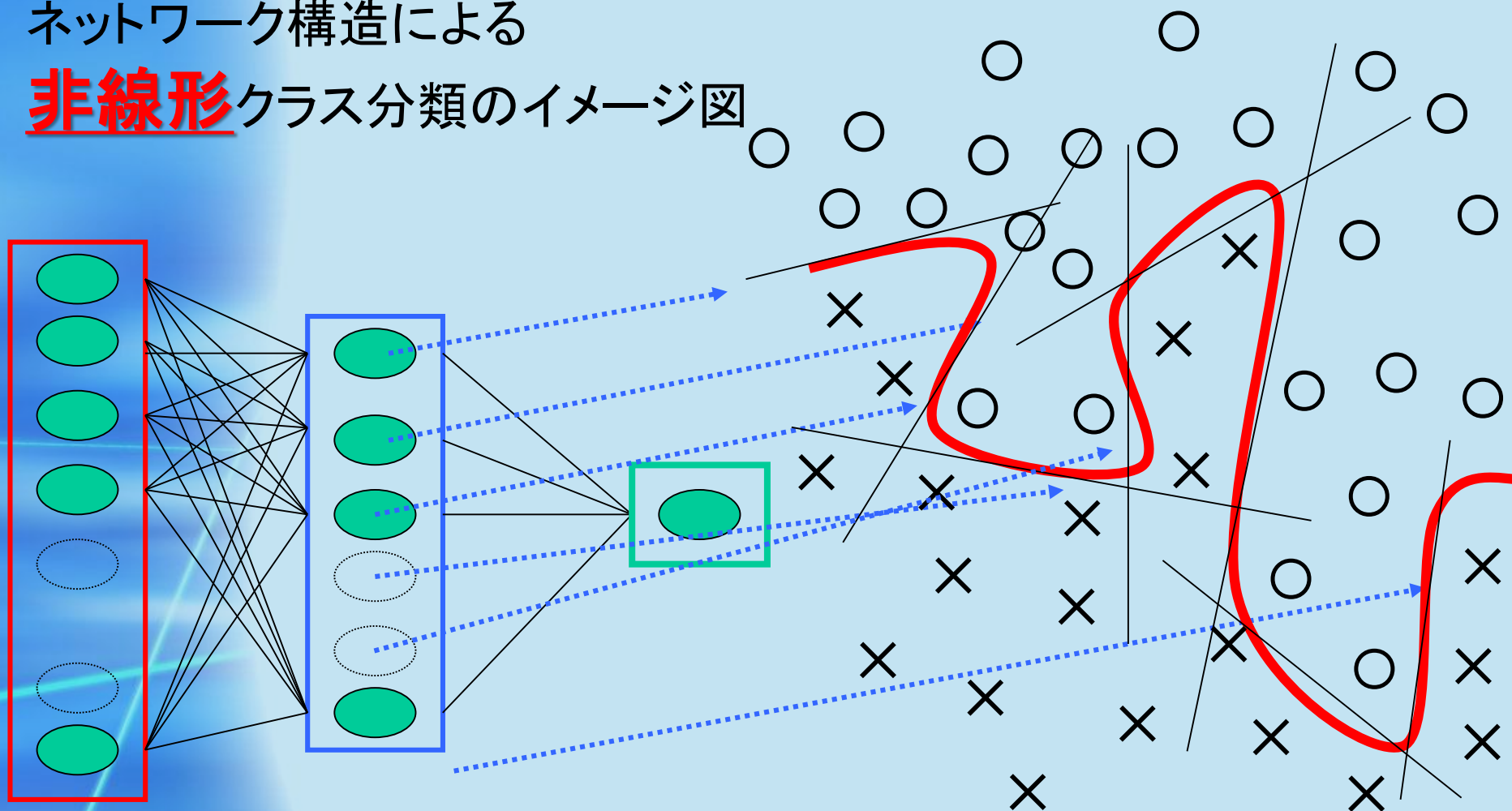


パーセプトロン
単純なネットワーク構造



ネットワーク構造による

非線形クラス分類のイメージ図

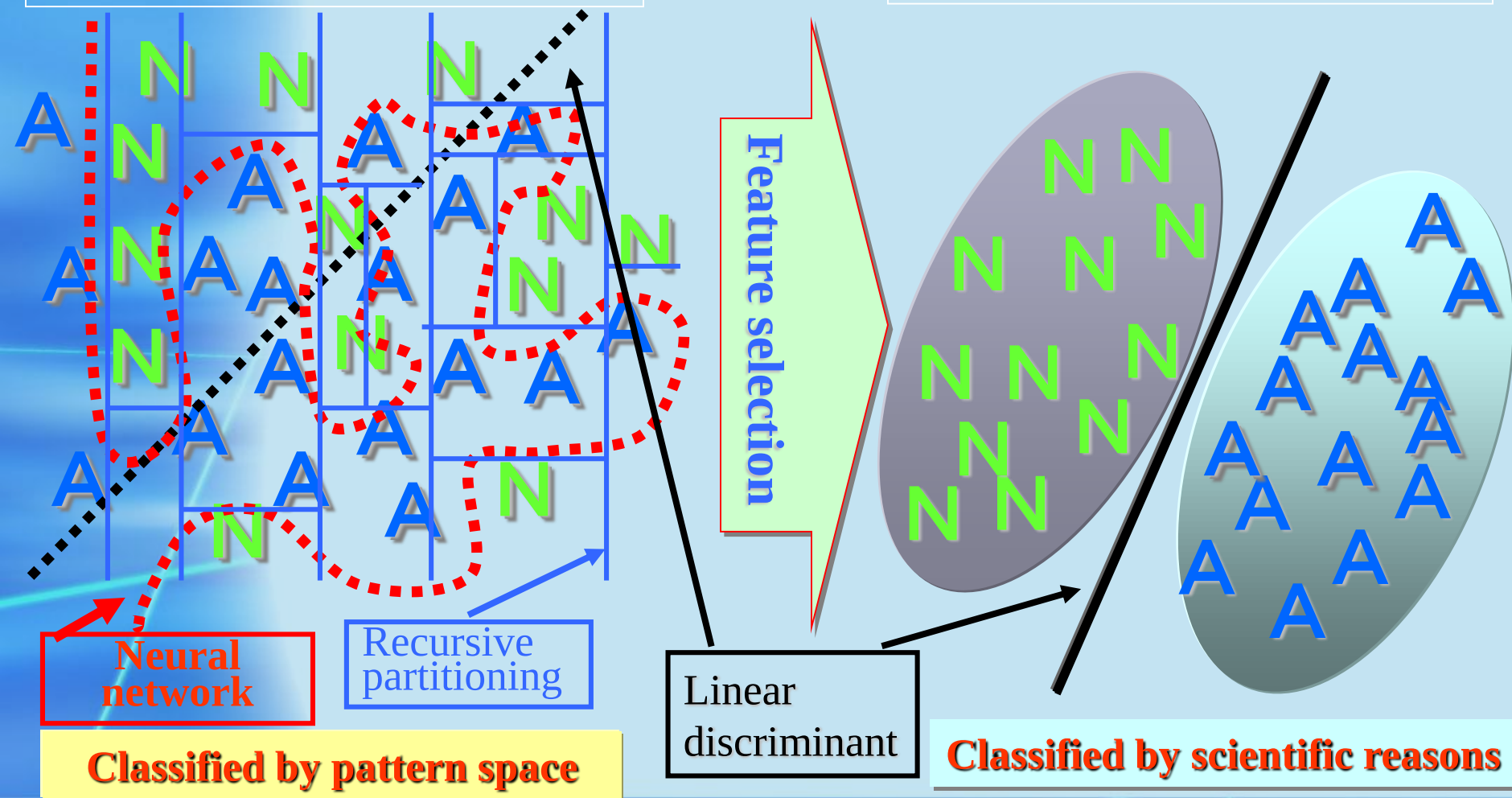


より複雑なネットワーク構造

Simple classification and scientific classification

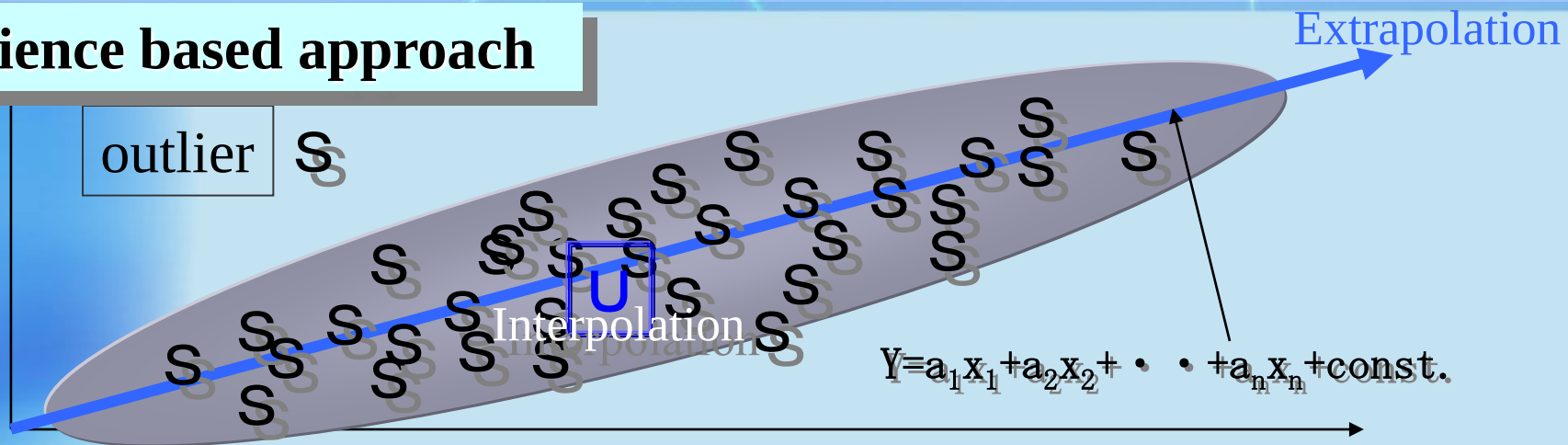
Pattern space impossible to be classified by linear discriminant

Pattern space classified by linear discriminant

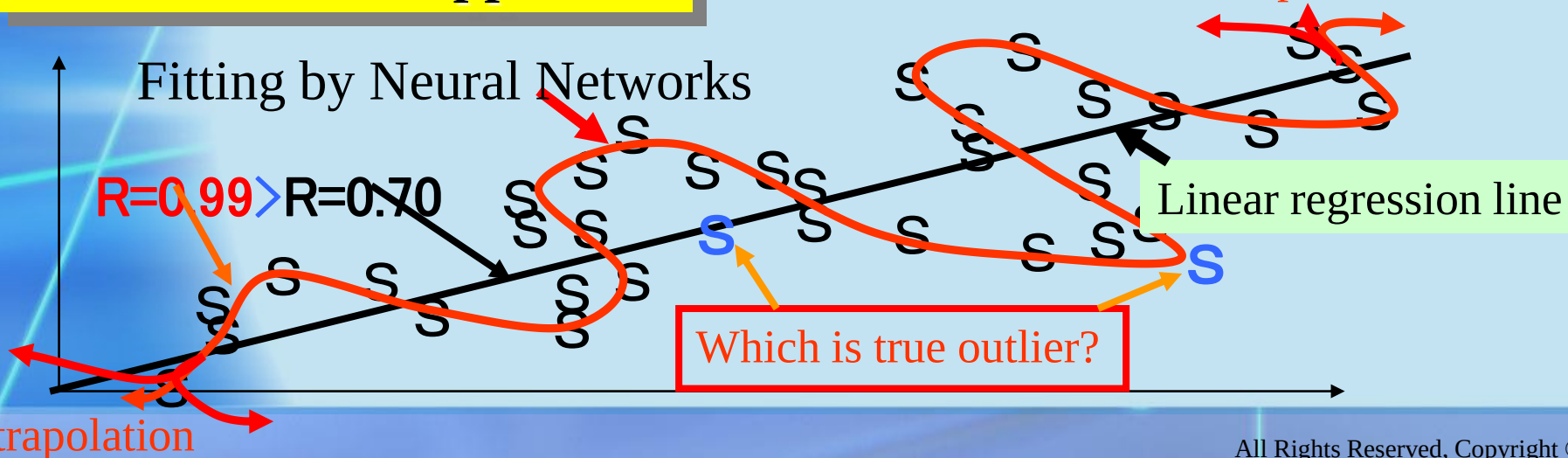


Simple fitting and scientific fitting

Science based approach



No science based approach



□ 線形解析の基本:

サンプル空間を作り上げるアプローチ

- ・ニクラス分類 $\Rightarrow$ クラス1とクラス2が交わらない空間を作る
- ・重回帰分析 $\Rightarrow$ サンプル群が回帰線上に集合する空間を作る

□ 非線形解析の基本:

**サンプル空間に判別関数や回帰式を合わせる
アプローチ**

- ・ニクラス分類 $\Rightarrow$ クラス1とクラス2が交わらない空間を作る
- ・重回帰分析 $\Rightarrow$ サンプル分布に合わせて回帰式を作る

□ 創薬関連手法の提案と実績

1. ケモメトリックスによる発がん性評価と種々要因解析
2. リード化合物検索⇒リード化合物スクリーニング
3. リード化合物再構築
⇒部分構造組情報を用いたリード化合物創出
4. SARDアプローチ⇒置換基効果による3D-QSAR
5. 人工知能とケモメトリックスによるマルチアクティビティ探索
⇒ドラグリポジショニングの実施
6. ケモメトリックス解析による「並列創薬」の提案
7. 「インシリコンビ」の提案⇒インシリコスクリーニング技術と
コンビナトリアルケミストリーの融合

□ ケモメトリックス(多変量解析／パターン認識) 関連提案と発明

1. 「超球法」の開発⇒最近隣法等への展開
2. 「強化削減法」の開発
⇒ニューラルネットワーク用のネットワーク簡素化手法
3. 「KY法」の発明⇒ニクラス分類KY法
4. 「KY法」の拡大展開⇒フィッティング(重回帰)への展開
ニクラス分類関連KY法:3種類開発
フィッティング(重回帰)関連KY法:3種類
5. 「テラーメードモデリング」の開発
6. KY法関連発明:2件追加
7. 深層学習関連発明:準備中

□ ケモメトリックス適用事例

人工知能システム(ルールベース型)とケモメトリックスによる 薬理活性予測機能を組み合わせた バーチャルドラグリポジショニングへの展開

■人工知能システム:EMIL

実施目的; 初期入力化合物の「生物学的等価体」を自動創出

- ①初期入力化合物の構造を、「生物学的等価」体に変換する
- ②変換するルールは、メディシナルケミストの構造変換ルール採用
- ③創出された化合物はバーチャル化合物である

◆ケモメトリックス: 活性／不活性のニクラス分類

実施目的; 複数の薬理活性に関する同時活性予測

- ①目的薬理活性(複数)に関する予測モデルを複数構築
- ②EMILにて創出された化合物を、①の判別関数で活性予測
- ③薬理活性有りの化合物は新規の薬理活性を有する化合物候補
- ④ドラグリポジショニング候補化合物の決定

人工知能とケモメトリックスによるドラグリポジショニング概念図

生物学的等価体
化合物創出

ドラグリポジショニングのチェック

EMILシステム

初期化合物
入力

化合物1
化合物2
.....
.....
化合物N-1
化合物N

判別関数 判別関数 判別関数 判別関数
1 1 1 1

○	×	×	×
×	×	○	×
..	..	..	..
..	..	..	..
×	×	×	×
○	×	×	○

(Example-Mediated-Innovation-for-Lead-Evolution)

EMILとは:

- ・京都大学教授(現名誉教授)発案による、ルールベース型人工知能システム

実施目的:

- ・メディシナルケミストが有する化合物変換ノウハウを 用いて
目的薬理活性向上を効率的に実現する化合物構造式変換を行う
人工知能システム
- * 変更された化合物は「バイオアナログウス(生物学的等価)」化合物と
命名された

変換ルールの取り出し:

- ・製薬および農薬関連企業における最先端のメディシナルケミストが召集され、
変換ルール取り出しチームを結成して実施

開発言語(当初):ハイブリッド形態

化合物関連操作等 ⇒ FORTRAN(科学関連分野言語)
ルール稼働用言語 ⇒ LISP(人工知能関連分野言語)

開発目的:

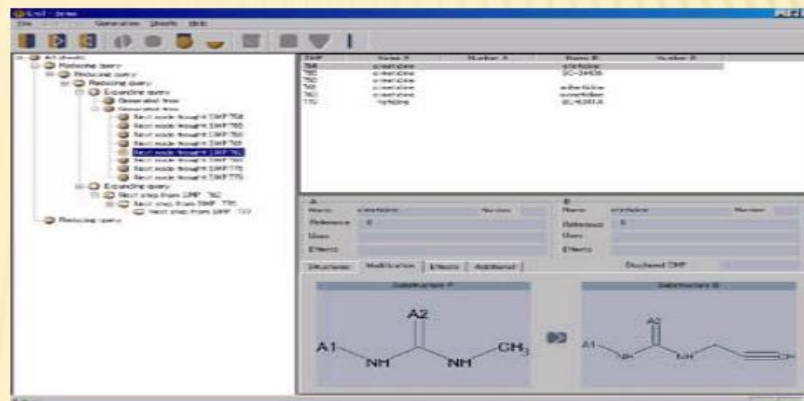
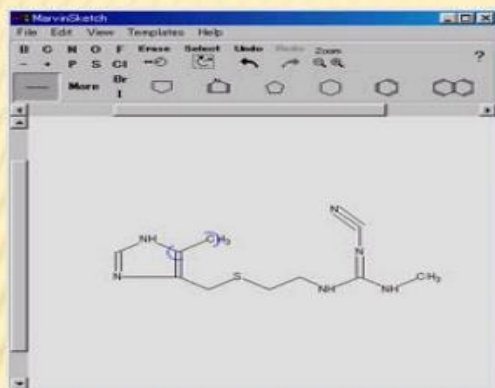
創薬研究者が有する化合物構造式の変換ノウハウ活用

- ・創薬研究者(特にメディシナルケミスト)は薬理活性の最適化を目指して化合物構造式のデザインが必要
- ・化合物構造式のデザイン可能性はほぼ無限にある
- ・開発目的に従った結果となるようにデザインしたい
- ・薬理活性のみならず、毒性や副作用回避も必要
- ・QSARやドッキングが適用困難な場合も多い

□プロの創薬研究者は、過去の豊富なデザイン経験より、効率的なデザイン手法についての様々な知識を持つ

貴重なノウハウ:活用したい

EMILシステムの流れ図



過去の構造変換ルールの適用
医薬品／農薬

バイオアナログウス
化合物群

化合物1

化合物2

化合物3

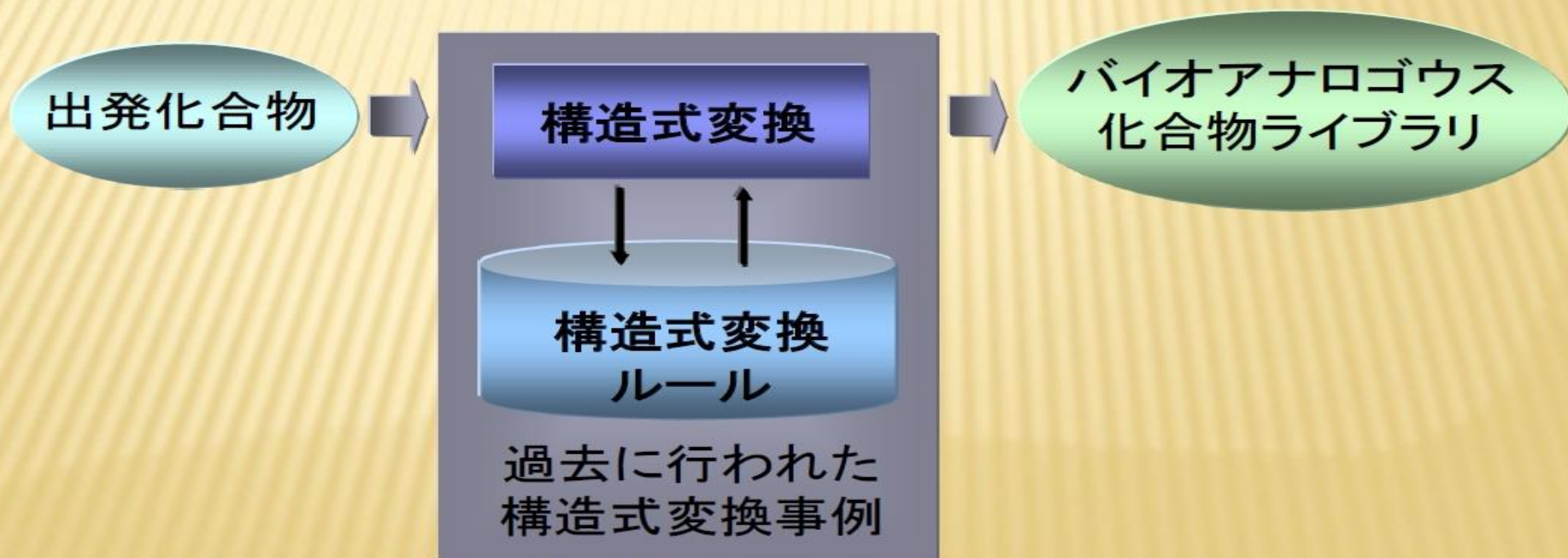
化合物4

化合物1-1

化合物1-2

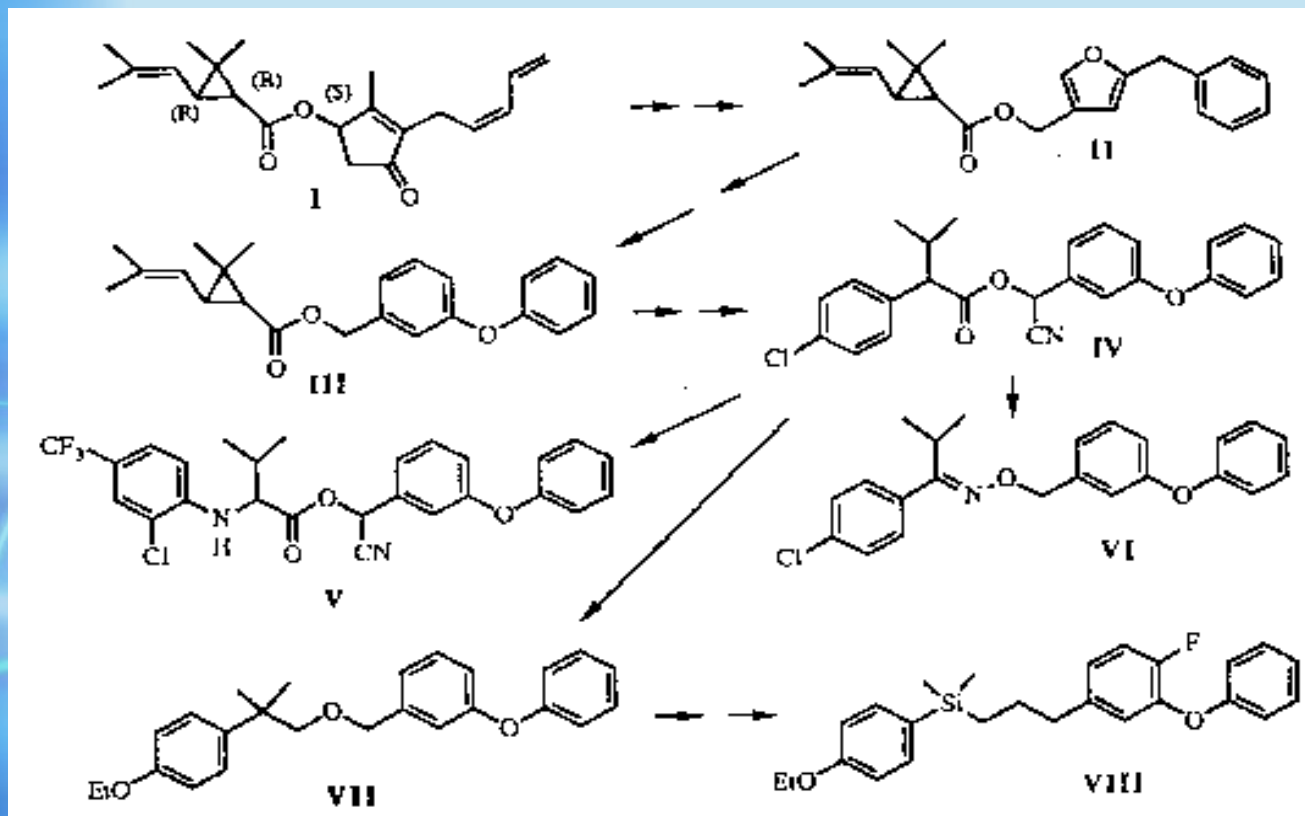
EMILシステム(人工知能システム)の機能と流れ

- 過去に行われた医薬品や農薬等の構造変換事例に従って、新たにバイオアナログウス化合物群を創出するプログラム



(Example-Mediated-Innovation-for-Lead-Evolution)

化合物構造のバイオアナログス変換事例



http://cicsj.chemistry.or.jp/14_1/14_1emil.htmlより引用

http://cicsj.chemistry.or.jp/14_1/14_1emil.htmlより引用

□ EMILシステムとケモメトリックスによる バーチャルドラグリポジショニング

実施目的:

既存薬物(Tiaramide)誘導体の新規薬理活性の探索

手順と基本技術:

1. 探索する新規薬理活性を5種類選択する
2. 薬理活性ごとに判別関数作成(高速スクリーニング用)
3. Tiaramideの誘導体をEMILを用いて生成
4. 生成された誘導体化合物について
ケモメトリックスによる薬理活性チェック

インシリコ創薬への人工知能の部分適用事例

マルチカテゴリースクリーニング実施事例

■ 5種類の薬理活性に関する同時スクリーニング

1. CCK-A (Cholecystinin-A) 受容体拮抗剤; Malonamide誘導体、急性膵炎治療薬
2. Ca^{2+} 遊離抑制作用; 1,5-benzothiazepine誘導体、降圧薬、抗狭心症薬
3. 抗潰瘍作用; 4-Bis[(aminoalkyl)amino]-9,10-anthracenediones誘導体
4. 抗菌作用; Quinolone誘導体
5. 抗アレルギー作用; Imidazo誘導体

■ 薬理活性スクリーニング対象化合物群 (化合物ライブラリー)

人工知能システムEMILを用いて創出された、バイオアナログウス化合物群

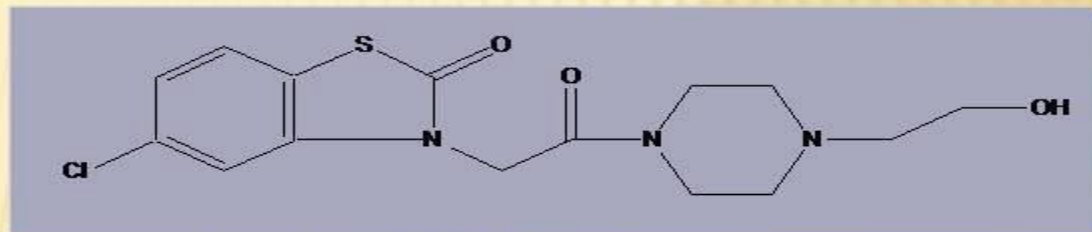
■ 薬理活性スクリーニング (多変量解析／パターン認識手法)

データ解析支援システムADAPTを用いて5種類の活性スクリーニングを行った

EMIL実行上での初期化合物と創出化合物群

■ 出発化合物構造式

Tiaramide
(Antiasthmatic, Anti-inflammatory)



■ 適用した構造式変換ルール

医薬品関連変換ルール

■ 変換結果

第一次変換出力構造式数 78化合物

第二次変換出力構造式数 76化合物

(第一次変換化合物中の1化合物を出発化合物とした)

化合物活性スクリーニング結果 (1)

1. CCK-A (Cholecystokinin-A) 受容体拮抗剤; Malonamide誘導体、急性膵炎治療薬
2. Ca^{2+} 遊離抑制作用; 1,5-benzothiazepine誘導体、降圧薬、抗狭心症薬
3. 抗潰瘍作用; 4-Bis[(aminoalkyl)amino]-9,10-anthracenediones誘導体
4. 抗菌作用; Quinolone誘導体
5. 抗アレルギー作用; Imidazo誘導体

化合物ID	活性1	活性2	活性3	活性4	活性5	活性総数
1	×	○	○	○	○	4
2	×	○	○	○	○	4
3	○	×	×	×	×	1
4	×	×	○	○	○	3
5	×	×	○	○	○	3
6	×	×	○	○	○	3
7	×	×	×	○	×	1
8	○	×	○	○	×	3
9	○	×	○	○	○	4
10	×	×	○	○	○	3
⋮						⋮

(Example-Mediated-Innovation-for-Lead-Evolution)

□ EMILシステムとケモメトリックス技術を用いた ドラグリポジショニング

実験結果:

新規薬理活性可能性ありとされる化合物を
効率よく探索できた

結果の考察:

1. 複数薬理活性候補化合物が頻度高く発見できた
EMILにより創出された段階で薬理活性候補化合物が
多数含まれた→EMILの構造変換ルールの効果を実証
2. 化学多変量解析／パターン認識(ケモメトリックス)により、
5種類の薬理活性について効率よくチェック出来た
3. 全てバーチャル化合物を用いて短時間で実施された

□ 解析原理・化合物操作・データ解析での留意事項

多変量解析／パターン認識によるデータ解析⇒宝箱・発見型アプローチ

化合物関連：一元一項対応、一次元表記、二次元表記、三次元表記

三次元の扱い問題：ローカル・グローバルミニマ対応

確率問題⇒偶然相関

フィッティング⇒オーバーフィッティング（過剰適合）

線形/非線形問題；ニクラス分類、フィッティング（重回帰）

特徴抽出（パラメーター手法）；種類、特徴、限界

サンプル関連；総サンプル数、最小サンプル数、

ポピュレーション（絶対数、クラス比率）

パラメーター関連：種類、単位の違い、オートスケーリング

ニューラルネットワーク⇒中間層の数 ⇒ 縦の重なり

・問題点；チャンスコリレーション、非線形性

パラメーター数が多いとき：パラメーター圧縮 PCA PLS

パラメーター数が少ない時；成功率低下、解析不能、物性式等

要因解析：パラメーターの読解力、分類力が強いパラメーター

KY法：サンプル数フリー、ポピュレーションフリー

ニクラス分類；完全分類

重回帰；高相関係数、高絶対係数

分類/予測⇒クロスバリデーション

1. ケモトリックスの適用範囲や研究分野は極めて広い
一つの研究分野が続くのはせいぜい20年～30年程度である
その後は新たな研究内容や手法への対応が必要となる
人生70年時代では、最低2種類の研究パターンへの対応が必要
ケモトリックスは適用範囲が広いので、時代の変化に対応可
 - ・ ケモトリックスの適用が出来れば、様々な研究活動可能
QSAR/QSTR/QSMR、 インシリコスクリーニング、他
 - ・ ケモトリックスを知れば、人工知能関連技術につながる
 - ・ 時代の変化や研究分野の変化に対応できる
 - ・ 他の研究分野での適用も可能
機器スペクトル関連分野; スペクトル解析、メタボロミクス解析
バイオ関連分野; 発現プロファイル解析、SNP's解析、
遺伝子解析、他
医学・医療関連; 法医学分野、画像解析⇒自動診断、他

まとめ:ドライブに例えて

2. 多変量解析／パターン認識の運用に関する知識がないと困る

- ・手法自体は多くの場合プログラム化されており、運用するだけ
車自体は多くの場合完成車が売られ、目的に合わせて選択
- ・運用を間違えると、解析そのものが全く意味が無くなる
運転技術を知らないと、事故を起こす
- ・多変量解析／パターン認識の適用を間違えると総てを失う
道路交通法を守らないと逮捕される(総てを失う)

3. データ解析結果の解釈と新規データ解析デザインの実施

- ・データ解析を行うのに、専門分野の知識が必要になる
車で目標地についてから行う作業は、その道の専門家が行う
- ・データ解析結果を受けて新たな実験を企画する
新たな目標地が決まれば、道路マップの作成と移動開始

Thank you for your attention

株式会社 インシリコデータ
湯田 浩太郎

<http://www.insilicodata.com>