◇本日の講義スケジュール

序章　１３：００～１３：２０
・自己紹介
・時代の流れと、研究環境および研究内容や研究スタイルの変化

第一章　１３：２０～１３：４０（１０分休憩）
化学データサイエンスおよび人工知能の適用イメージ

第二章　１３：５０～１５：００（１０分休憩）
化学データサイエンスおよび人工知能の適用事例：様々なアプローチが可能

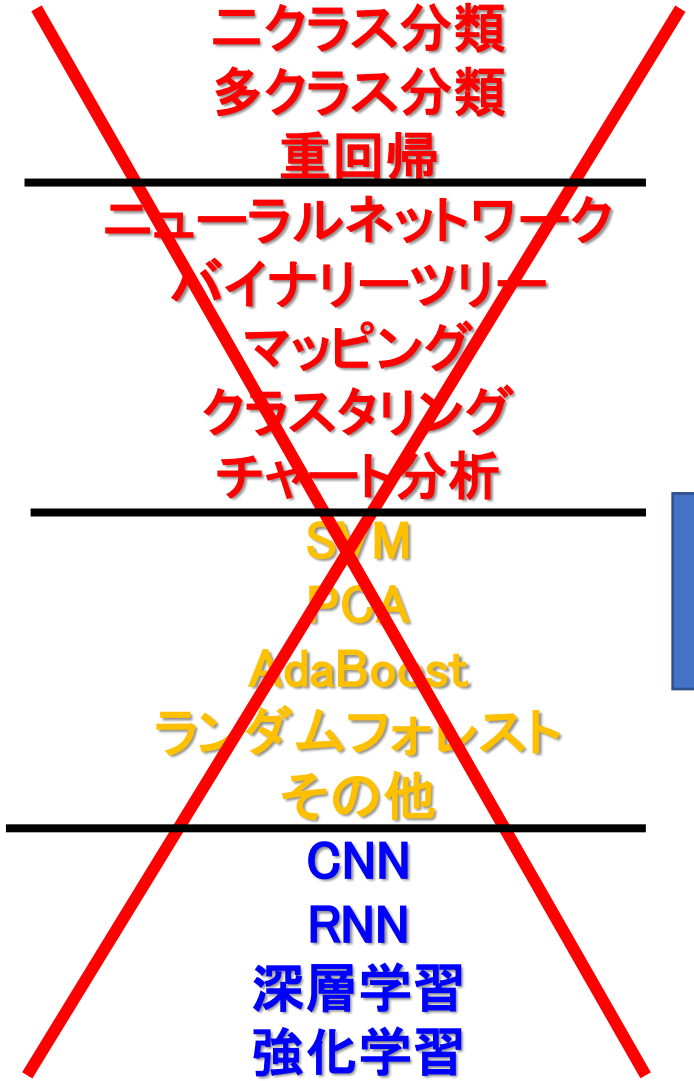
第三章　１５：１０～１５：５０（１０分休憩）
化学データサイエンスおよび人工知能の適用手順やパターン

第四章　１６：００～１６：５０
化学データサイエンスおよび人工知能の適用上での留意事項
　①化学分野の留意点
　②データサイエンス実施上での留意点

まとめと提案　１６：５０～１７：００
　①化学データサイエンスおよび人工知能のまとめ
　②今後の展開についての「オートノマス創薬」の提案
　・KY法の展開（クラス分類と重回帰型）
　③自由討論

チュートリアルの主要な内容

データ解析手法 基本原理等



基本的な部分は概略説明

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
○その他

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

□本日の討論内容

In Silico Data
Miracles by the KY-methods

□DS（データサイエンス）と人工知能適用における化学的問題

□機械学習における問題

□偶然性（チャンスコリレーション）の問題

□過剰適合/過学習

□機械学習型人工知能

□化合物観連一元一項対応の重要性

□DSと人工知能適用における化学的問題

創薬は化合物構造式中心の世界

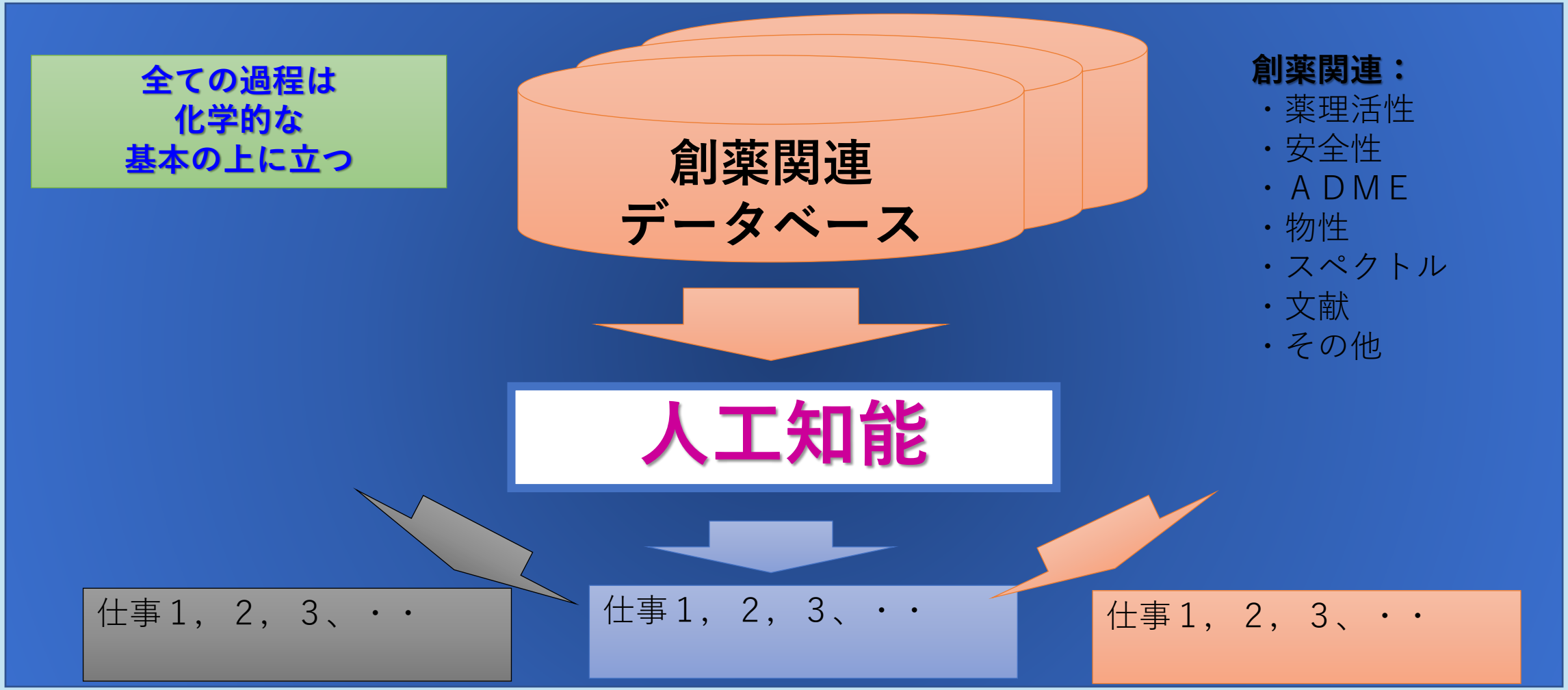
化学研究者の思考過程は化合物構造式で考え、
相互コミニュケートし、化合物構造式で答える。

人工知能システムが、利用者である研究者と、
化合物構造式で対話できることが重要

例：創薬研究者
薬理活性を強くするには、化合物構造式のどの部分を
どのように変化させればいいの？⇒研究者との対話必要

チェス、将棋、碁のように、盤上の座標を指定するようにはゆかない
また、勝つだけで良いというわけでもない

□DSと人工知能適用における化学的問題



□DSと人工知能適用における化学的問題

□化合物構造式に始まり、化合物構造式に終わる

・研究者の思考過程は総て化合物構造式で終始する（特にメディシナルケミスト）

・**化合物の表現の問題：**

化合物名、分子式、二次元構造式、３次元構造式、等々
同じ化合物が表現系により様々な形式を取り、それぞれの
表現系が持つ情報の内容や情報量も異なる。

・**入力の問題：**

Ｊｏｕｒｎａｌ　や一般の化学文献が膨大な量あっても、人工知能の学習に必要となる化合物構造情報を正確に入力させることが必要。

・**結果の問題：**

結果が出たら、人工知能情報の化学情報への変換が重要

□DSと人工知能適用における化学的問題

In Silico Data
Miracles by the KY-methods

例：化合物の「一元多項」問題

- 人工知能に複数の顔で入ってきた化合物の扱い？
  同一化合物であることをチェックする機能必須
- 学習過程で異なる化合物と判定される可能性

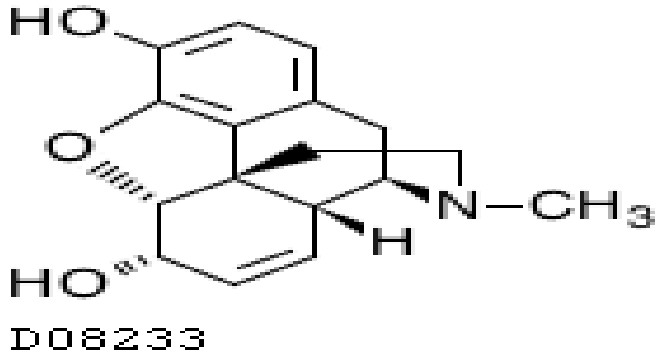
例：Journal情報利用上での問題

- 化学やバイオ関連分野の論文は基本的に成功事例
  成功のみ掲載されている。このような成功事例のみを
  学習した結果提案される化合物は、
  成功／失敗化合物？　→　失敗というフィルターがない
- 入力Journal数は精度の保証にならない
  数が多いほど上記の偏向学習が進んでいることの証拠

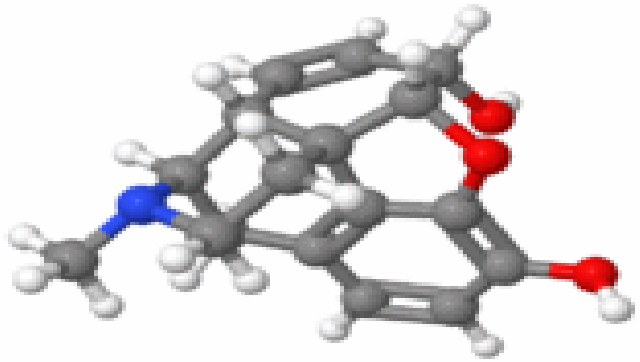
□DSと人工知能適用における化学的問題

In Silico Data
Miracles by the KY-methods

化学者の思考過程で必須

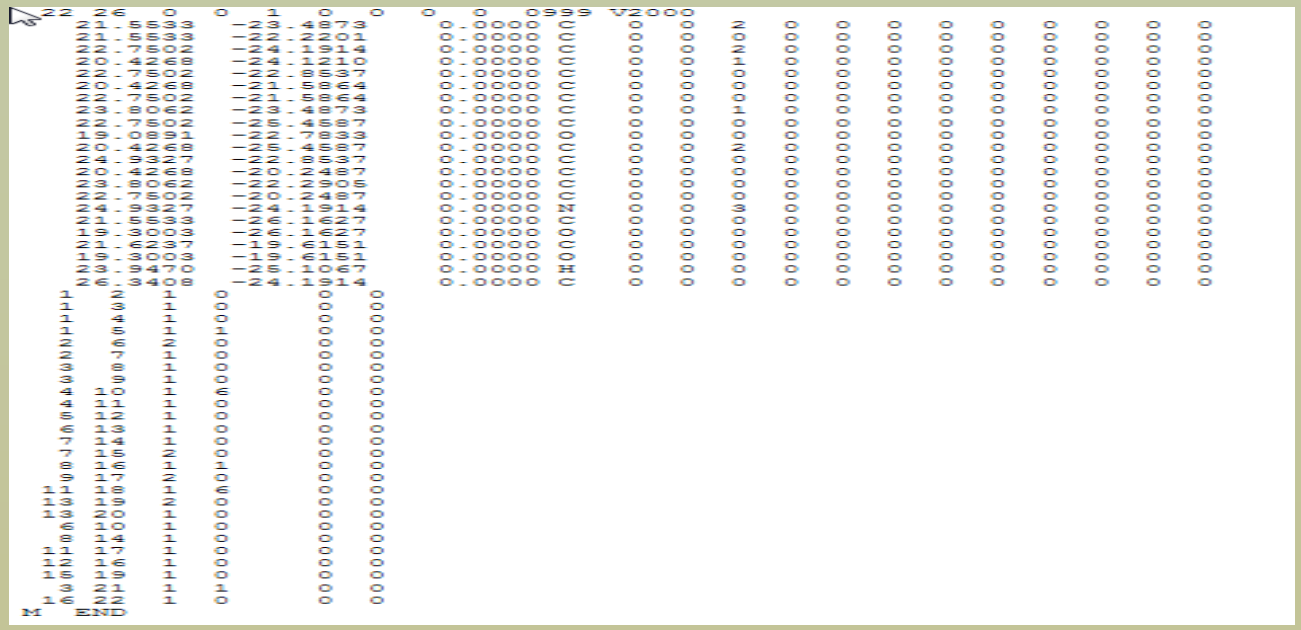


D08233



化学者はイメージも議論も出来ない

Ｍｏｌファイル：



IUPAC名：

(5α,6α)-7,8-didehydro-4,5-epoxy-17-methylmorphinan-3,6-diol

ＳＭＩＬＥＳ名：

OC(C=CC1CC2N3C)=C(OC4C(O)C=5)C1C4(CC3)C2C5

□DSと人工知能適用における化学的問題

＊化学者がイメージできる情報は化合物構造式で、
　数字や文字だけでは議論も出来ない

＊コンピュータが扱えるのは数字と文字コードで、
　構造式のイメージ情報は扱えない

人工知能実施の上で、上記２事象間の
ギャップを埋めることが必要

□機械学習における問題

In Silico Data
Miracles by the KY-methods

□最近の人工知能は機械学習がメインである

利点：

・大量のデータを扱える
・従来は人工知能で展開出来なかった内容を展開できる
・ノウハウ（ルール）等を必要としない：データがあれば良い
ノウハウがない分野での展開が可能となる
・**新たな知見**を発見出来る可能性がある

欠点：問題点

・化学的な知見をシステムに理解させられるか？
・結果のフィードバックが手法的に困難
・**新たな知見**を人間が解釈できるレベルへの具象化が困難

□機械学習における問題

In Silico Data
Miracles by the KY-methods

□深層学習実施上での留意点

1．学習に用いる**サンプル数問題**→過剰適合の回避
　ネットワークの階層が深いと、学習に必要なサンプル数が
　急激に増大
　現時点（2016.11.13）での**ＣＡＳ**登録化合物（有機／無機）数
　⇒123,623,093
2．**学習の偏り**回避→サンプルの学習内容が大事
3．現在は**画像／音声／文字認識が主体**のネットワーク構成
4．結果が良くても、**ネットワークから要因情報を取り出す**ことが
　極めて困難

□サンプル数に関係する問題

1．総サンプル数

信頼性の高い解析をするにはある程度サンプル数が多いことが必要である

・データ解析を行うにあたって、最小サンプル数は？と問われることが多い。
ニクラス分るや重回帰にはサンプル数とパラメータ数との間で信頼係数が設定される。
この、解析信頼性を基準に１パラメータ用いた時、**サンプル数はニクラス分類で４サンプル**、**重回帰では６サンプル**必要である。
・必要なサンプル数はデータサイエンス手法により異なる。複雑なネットワーク構造を有するニューラルネットワークや深層学習では解析に必要なサンプル数は飛躍的に増大する。

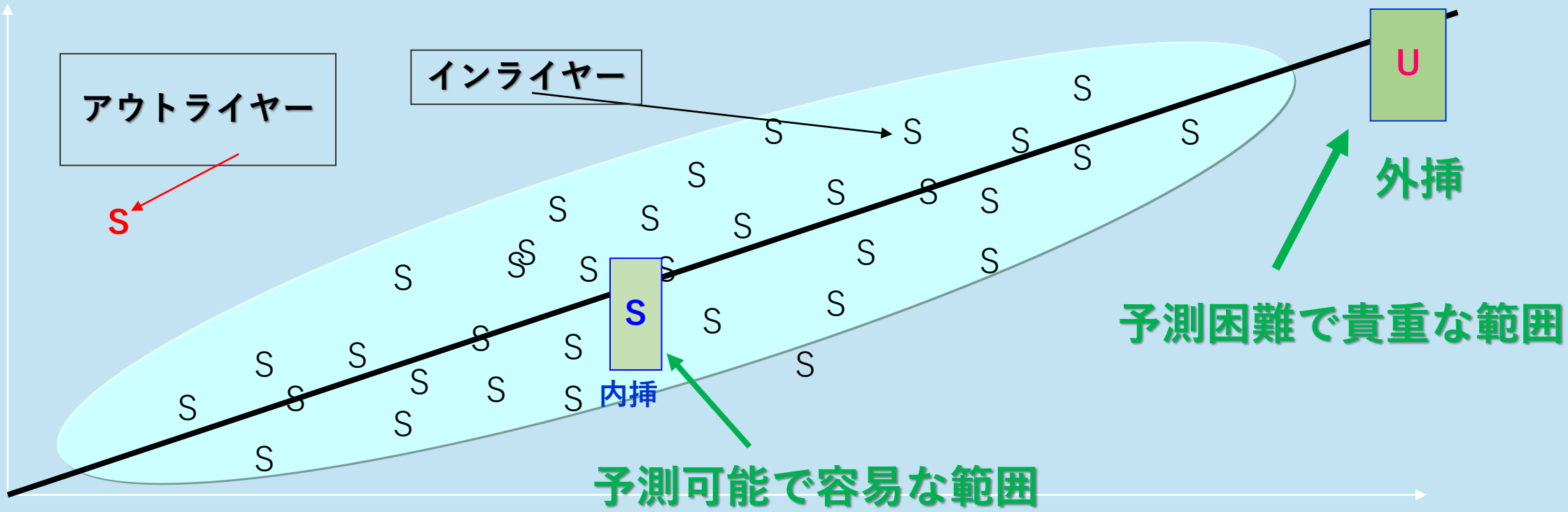
2．クラスポピュレーション（サンプル数にクラス間の偏りは禁じ手）

クラスを構成するサンプル数も適切な配分が必要である

・ニクラス分類では、小さなサンプル数のクラスは、**分類に利用するパラメータ数と同じ数**のサンプル数が必要となる
・ニクラス分類で、サンプル比が１０：１程度以下になると分類の精度は**サンプル数の多いクラスに有利**になる。サンプル比がもっと低くなると、サンプルの少ないクラスは総て無視されて、サンプル数の多いクラスに分類される。即ち、小さいクラスは１００％誤分類となる。それでも、全体の分類率は高くなる。

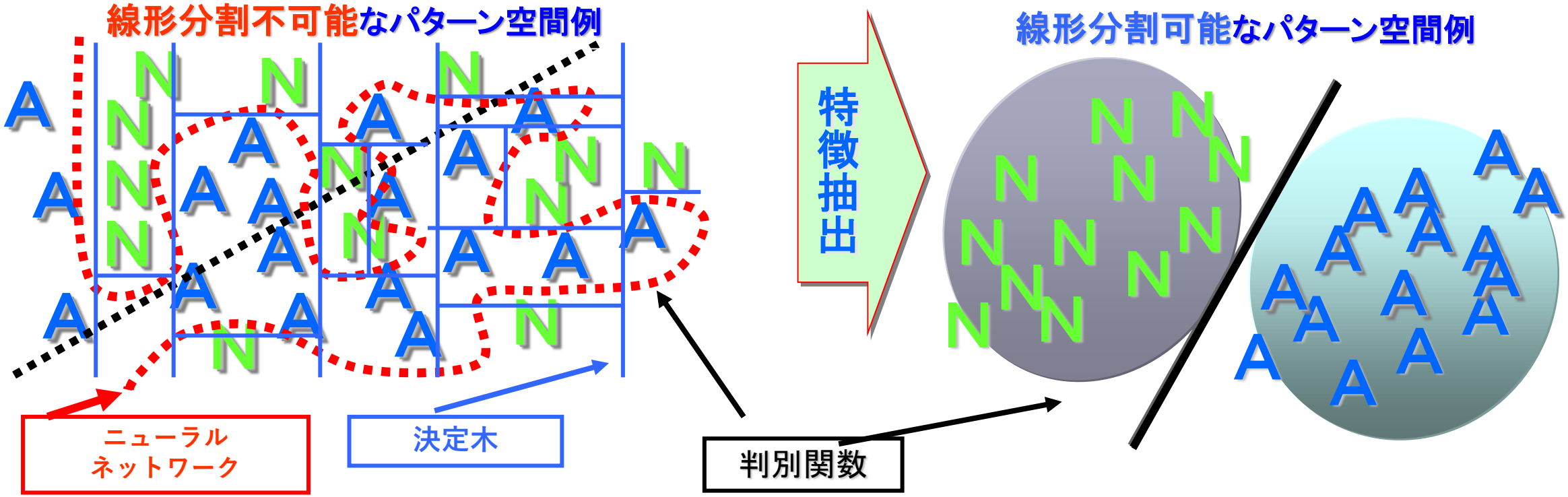
□機械学習における問題

線形重回帰におけるアウトライヤーとインライヤー
内挿性と外挿性



□機械学習における問題

ニクラス分類問題における過剰適合問題



過剰適合発生状態での分類

空間にＡとＮを分ける情報を持たない

空間を形成するパラメータ群は
ＡとＮの識別根拠情報が見えにくい

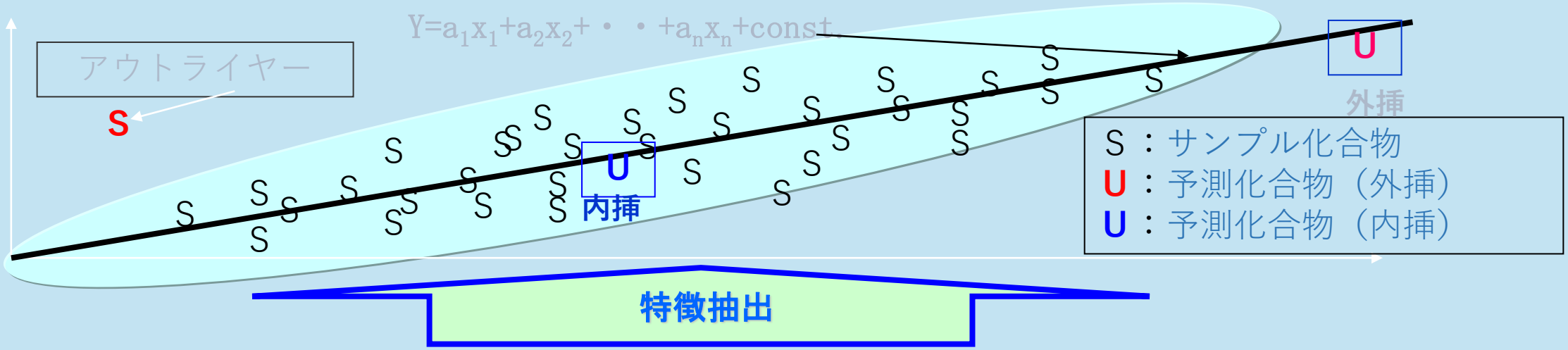
分類目的を達成する空間構築による分類

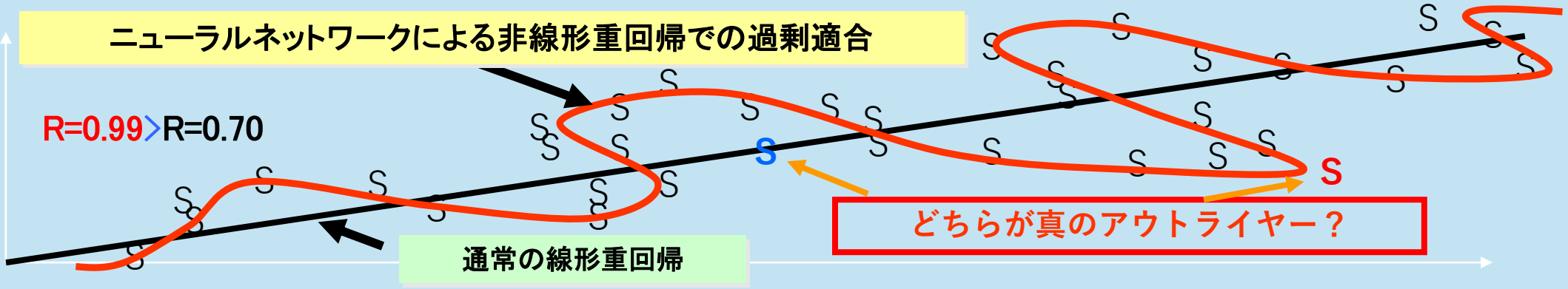
空間にＡとＮを分ける情報を持つ

空間を形成するパラメータ群は
ＡとＮの識別根拠情報が見えやすい

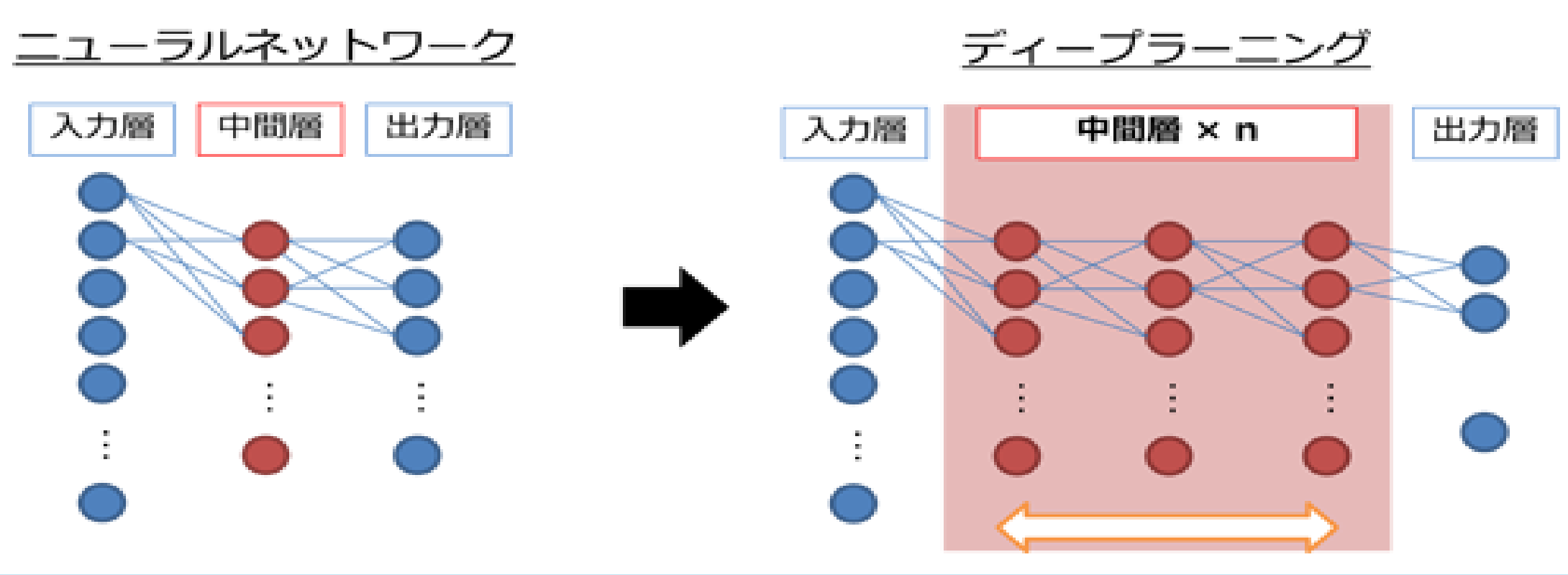
□機械学習における問題

線形重回帰における過剰適合問題





□機械学習における問題



多層化による勾配消失と過剰適合の問題があったが、近年、アルゴリズムの改良とデータ量の増大、そして膨大なデータを処理できる計算装置（コンピュータ）の爆発的な性能向上によって問題が解消されてきた。

https://markezine.jp/article/detail/24185より抜粋

□生命科学分野における今後の人工知能の展開

In Silico Data
Miracles by the KY-methods

□展開分野

歴史的に化学生物分野での人工知能展開事例は多い

□実施手法

現実の条件に即した機械学習の改良／開発

□ハイブリッド型

・機械学習およびルールベース
・多変量解析／パターン認識との連携

□ｌｏｔや分析／医療機器との連携

□その他の展開

□今後の人工知能の展開

In Silico Data
Miracles by the KY-methods

□大量データの効率的な使用が可能か

・データの形式や、情報内容が不揃いである場合は機械学習で扱うことが困難であり、人工知能用にデータの整理が重要
・量があっても、人工知能の実施目的に必要な情報が取り出せない、あるいは偏った情報では学習に使えない

□化学構造式の正確な理解を機械学習で行う技術

・化学特有の様々な問題を、大量のデータから自動的に学習することには困難が予想される

創薬や安全性等の化合物を解析対象とする場合、機械学習のみならず、既存のノウハウ導入や、多変量解析/パターン認識（ケモメトリックス）技術等との連携を念頭に、総合的なアプローチを考えるのが最も合理的である

□偶然性（チャンスコリレーション）の問題

　データ解析を実施する時は、偶然に好ましい結果が出ることがあり、これを回避することが信頼性の高いデータ解析の実施に繋がる。従って、この回避指標はデータ解析実施上での極めて重要な指標である。
　この回避指標はデータ解析手法ごとに異なる指標を有するので、注意が必要である。

線形重回帰：チャンスコリレーションの回避指標

$$| = \frac{\text{サンプル数}}{\text{利用パラメータ数}} \geqq 6$$

この値が６以下の値の時、その解析結果の信頼性は無いので解析結果は採用されない

□偶然性（チャンスコリレーション）の問題

ニクラス分類の場合：以下の条件を満たすことが必要

$$I = \frac{\text{サンプル数}}{\text{パラメータ数}} \geqq 4$$

◇偶然性の考え方

通常起こりえない事が起こる

↓

必然性：相関

ニクラス分類の時、１パラメータで１００サンプルが正解する時、この１パラメータは分類するに必要な相関情報を有すると言える。

必ず起こる事が起こる

↓

偶然性：無相関

ニクラス分類の時、10000パラメータで100サンプルが正解しても、この10000パラメータは分類するに必要な相関情報を有しないと言える。

■本日のプログラム

◇ケモメトリックス解析を保証するための最低限の制限事項

□ 「偶然性」問題における次元数とサンプル数との関係（一般化）
次元（記述子）が一つ増える毎に分類可能な場合の数は２倍ずつ増加する。
従って、次元数ｄにより定まる分割可能な場合の数Ｒは以下の式で示される。

$$R = 2^{d} \qquad (1)$$

この結果、次元数がＮで、サンプル数が2^{N}以下の時には必ず分類出来、この分類結果は偶然により支配されている事は明白である。
一方、サンプル数がｎの時、このサンプルを２クラスに分類出来る場合の数Ｃは単なる組み合わせ問題であり、以下の式で示される。

$$C = \frac{1}{2} \sum_{k=1}^{n} \frac{n!}{k \times (n-k)!}$$

これらの項目を考慮し、与えられた記述子（次元数）ｄでサンプルｎを２分割出来る可能性Ｐは

$$P = \frac{\text{サンプルｎに対する２分割の場合の数}}{\text{記述子ｄによる２分割の場合の数}} = \frac{C}{R}$$

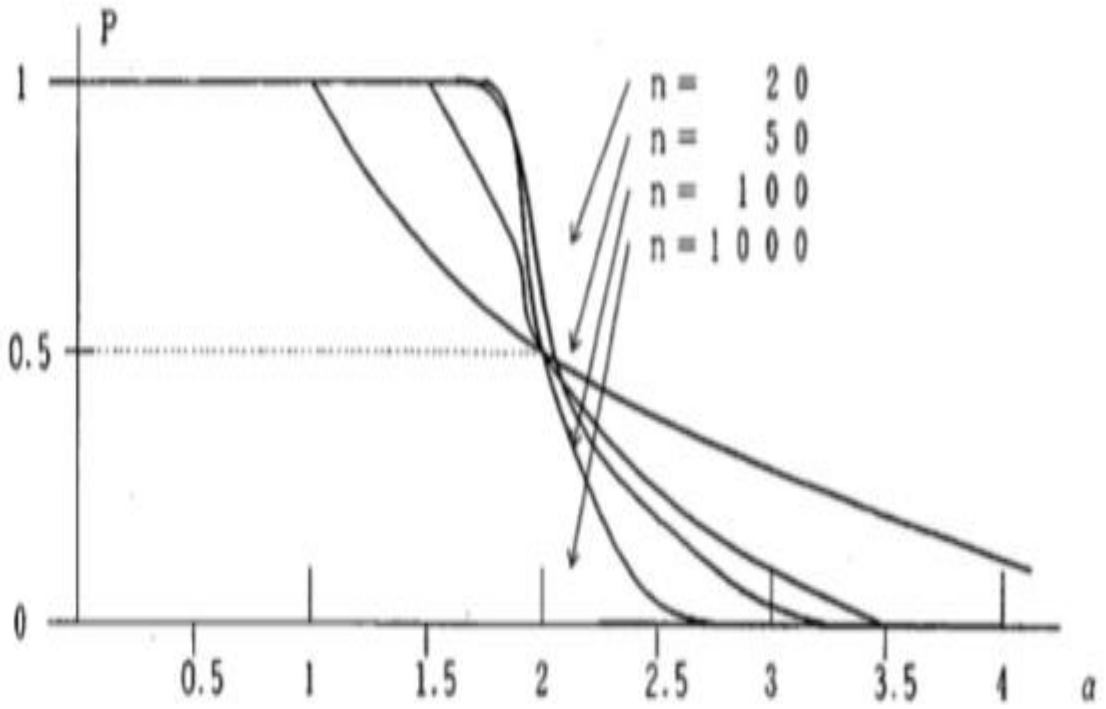
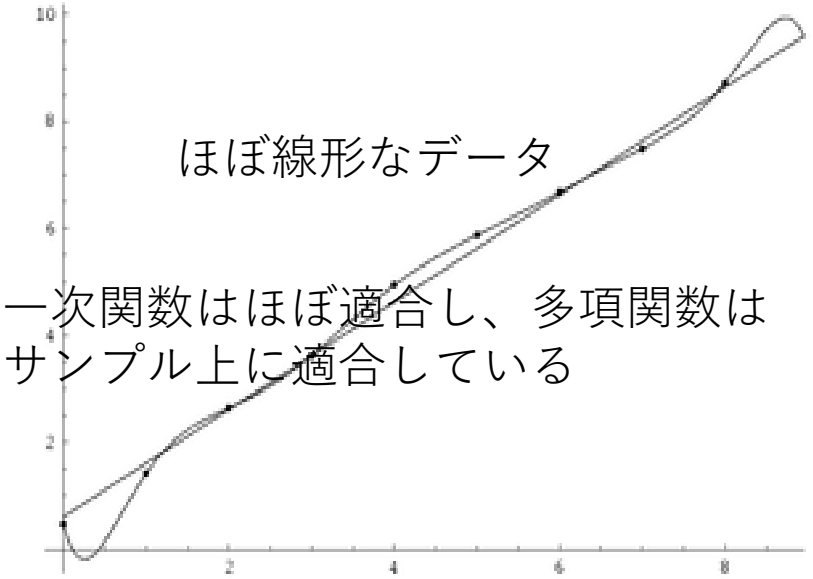


図４．　２分割の可能性に対するα（サンプル数ｎ／次元数ｄ）とＰの関係

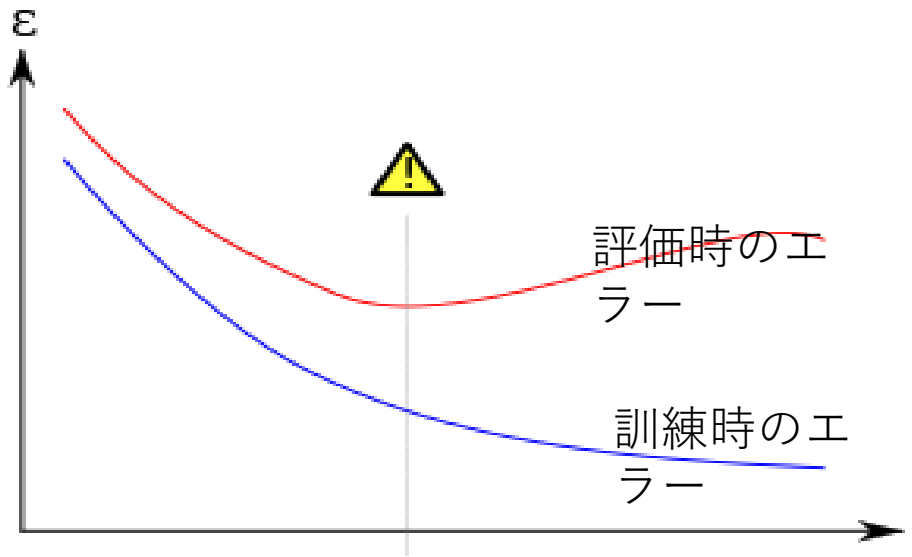
□過剰適合/過学習

過剰適合（かじょうてきごう、英: Overfitting）とは、統計学や機械学習において、訓練データに対して学習されているが、未知データ（テストデータ）に対しては適合できていない、汎化できていない状態を指す。汎化能力の不足に起因する。



一次関数はほぼ適合し、多項関数はサンプル上に適合している

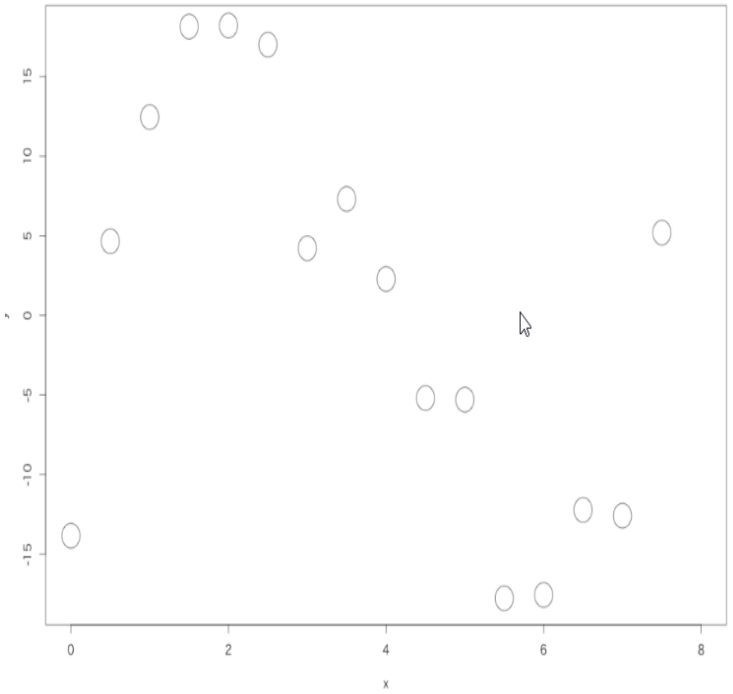
一次関数は両端で値が安定するが多項関数は両端で値が大きく変動



ニューラルネットワークでの過剰適合の状況

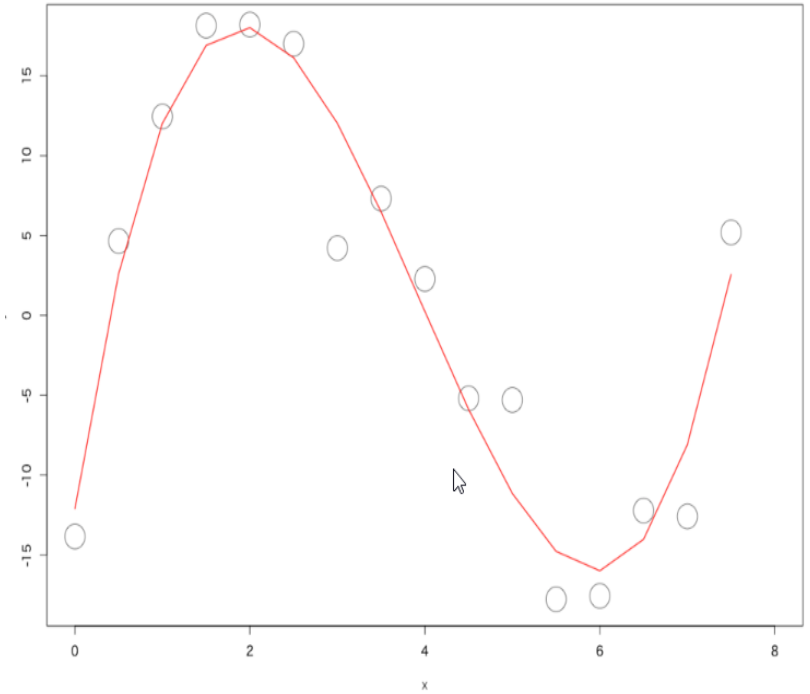
https://ja.wikipedia.org/wiki/過剰適合

□過剰適合/過学習



全サンプル数２１
学習用１６サンプル

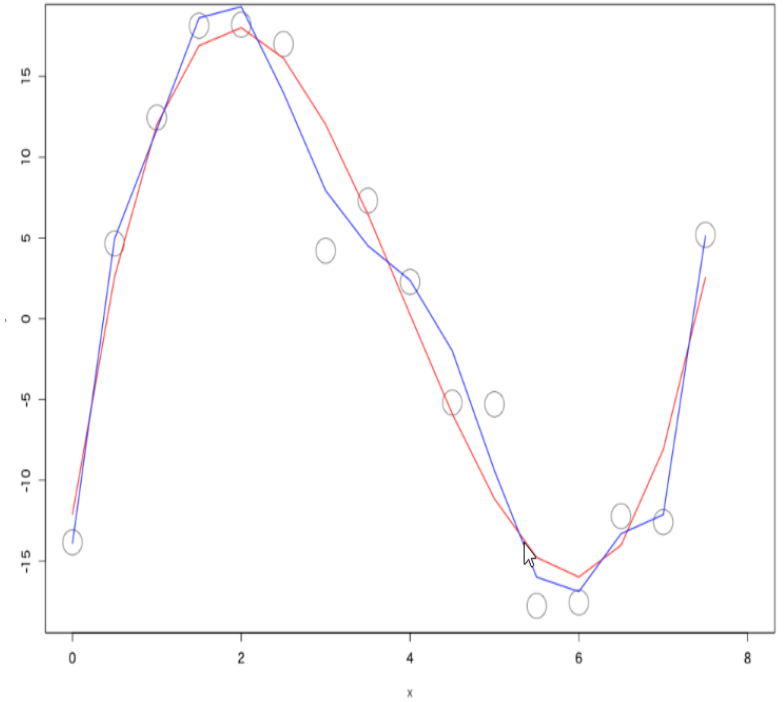
３パラメーターを用いた重回帰
解析信頼性を保ったパラメーター数



相関係数：　０．９６８

汎化能大

９パラメーターを用いた重回帰
解析信頼性を伴わないパラメーター数

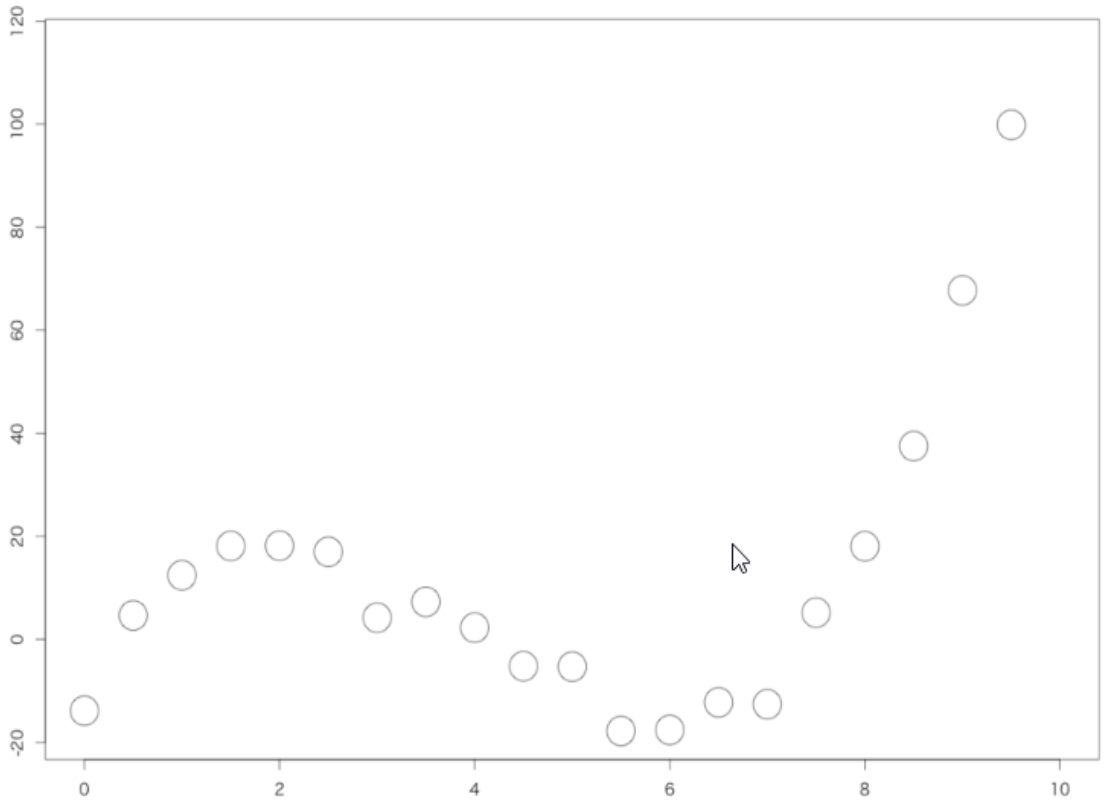


相関係数：　０．９８６

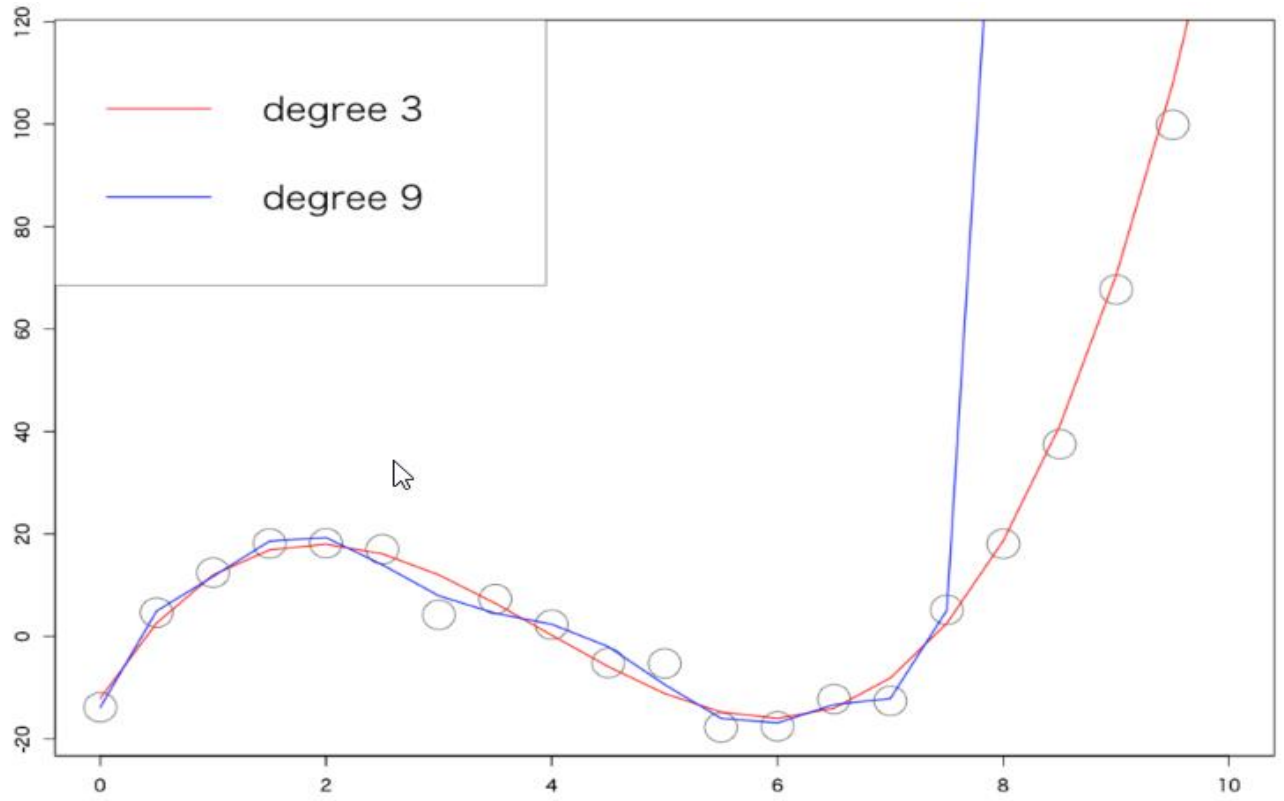
過学習

https://tjo.hatenablog.com/entry/2016/04/14/190000

□過剰適合/過学習



全サンプル２１
学習用１６＋テストデータ５

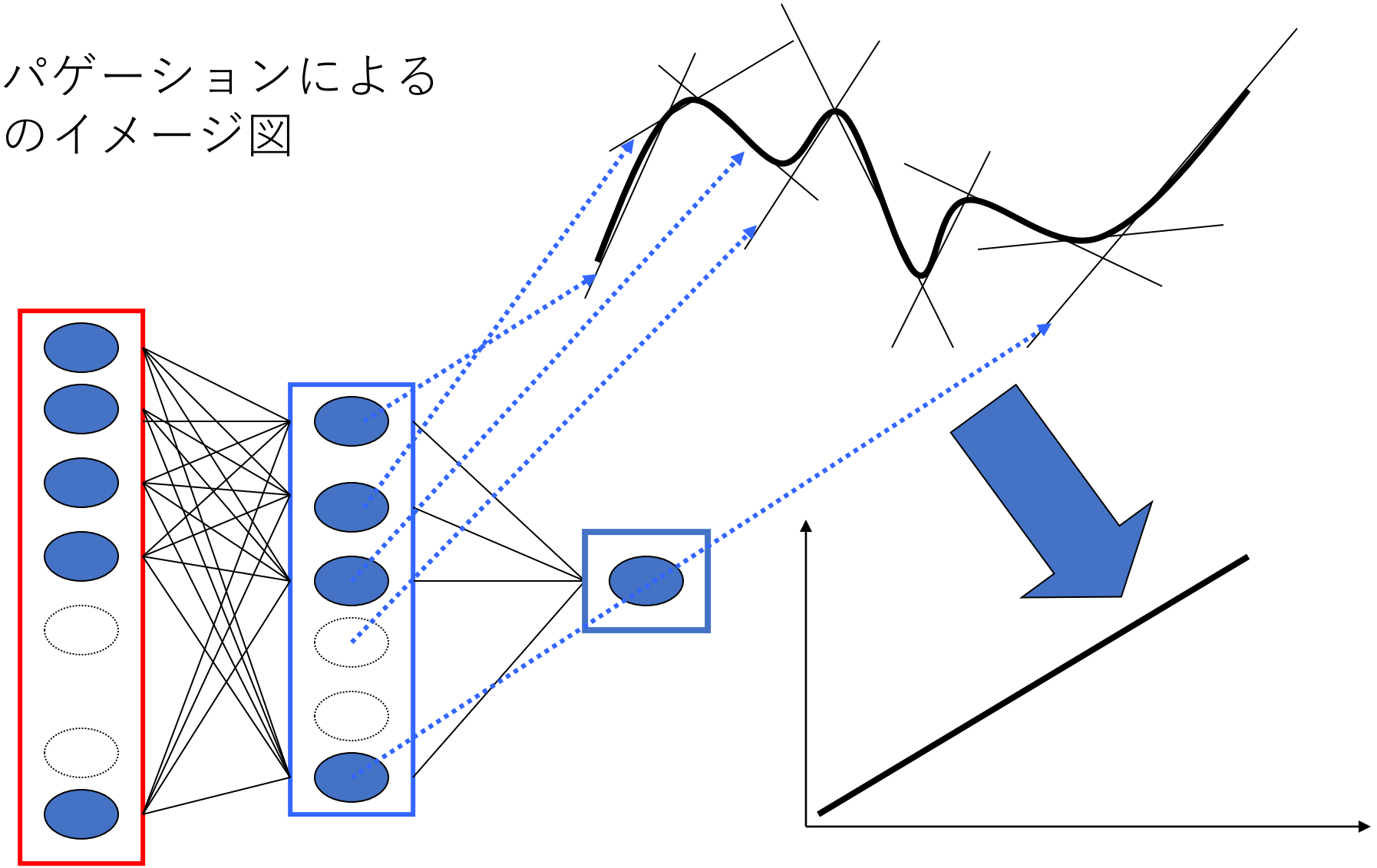


パラメーター３の場合と９の場合の回帰図

https://tjo.hatenablog.com/entry/2016/04/14/190000

ニューラルネットワーク：バックプロパゲーション

バックプロパゲーションによる
非線形分類のイメージ図

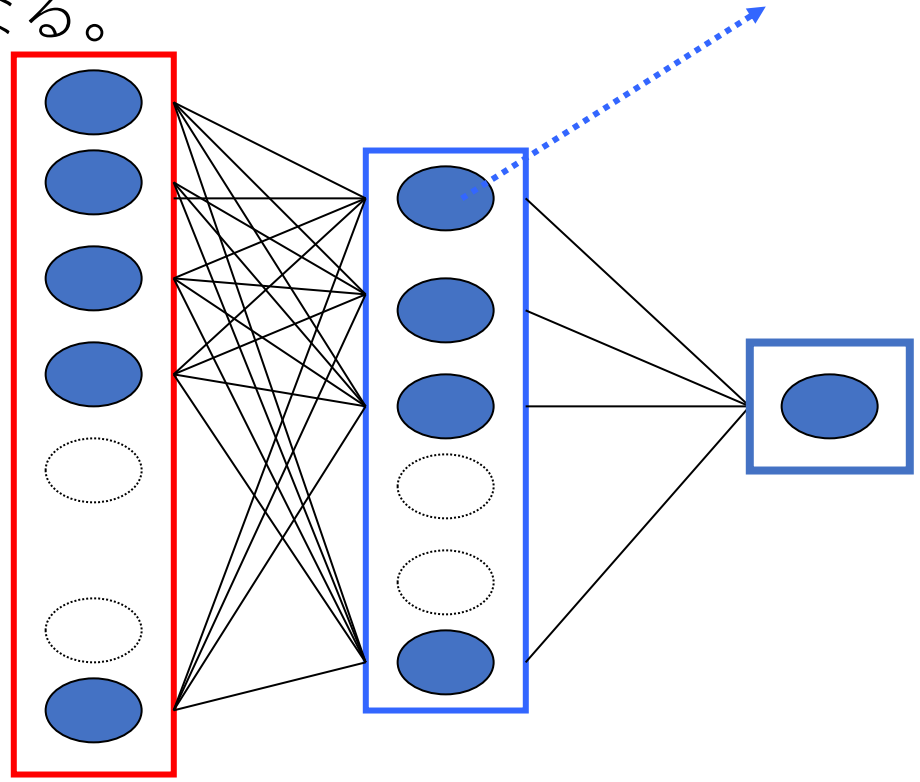


□ニューラルネットワークでのニクラス分類問題

・ニューラルネットワークは究極の非線形分類手法となる。
　従って、ニューラルネットワークでは利用するパラメータ数は少なくすることが偶然相関を避ける意味で重要である。
・ニューラルネットワークでは中間層のユニット数もハイパーパラメータとして指定できるが、この中間層のユニット数を大きくすると、パラメータ数と関係なく偶然相関の可能性を上げることになる。

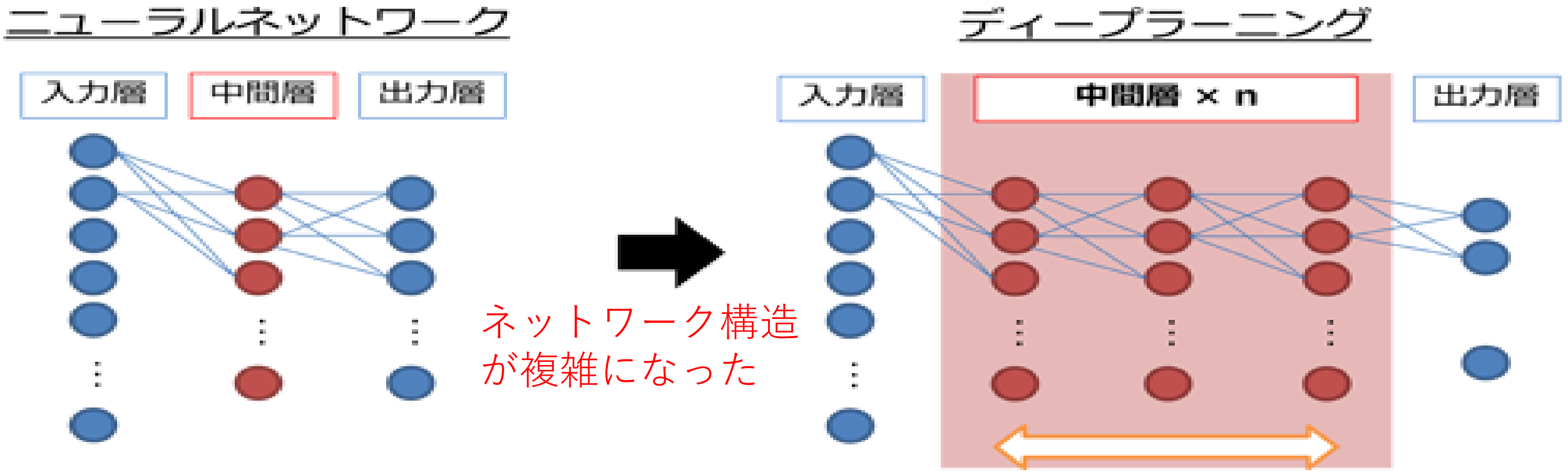
1．入力層のパラメータ数を少なくする
2．中間層のユニット数も少なくする

ニューラルネットワークであると、通常のニクラス分類よりも高い分類率を獲得しやすくなると一般的に言われるが、これはニューラルネットワークのネットワーク構成を考えれば当然の帰結である。



□ニューラルネットワークから深層学習に

・機械学習における問題



多層化による勾配消失と過剰適合の問題があったが、近年、アルゴリズムの改良とデータ量の増大、そして膨大なデータを処理できる計算装置（コンピュータ）の爆発的な性能向上によって問題が解消されてきた。

https://markezine.jp/article/detail/24185より抜粋

□深層学習でのネットワーク構造と留意点

・機械学習型人工知能適用上での解決すべき点

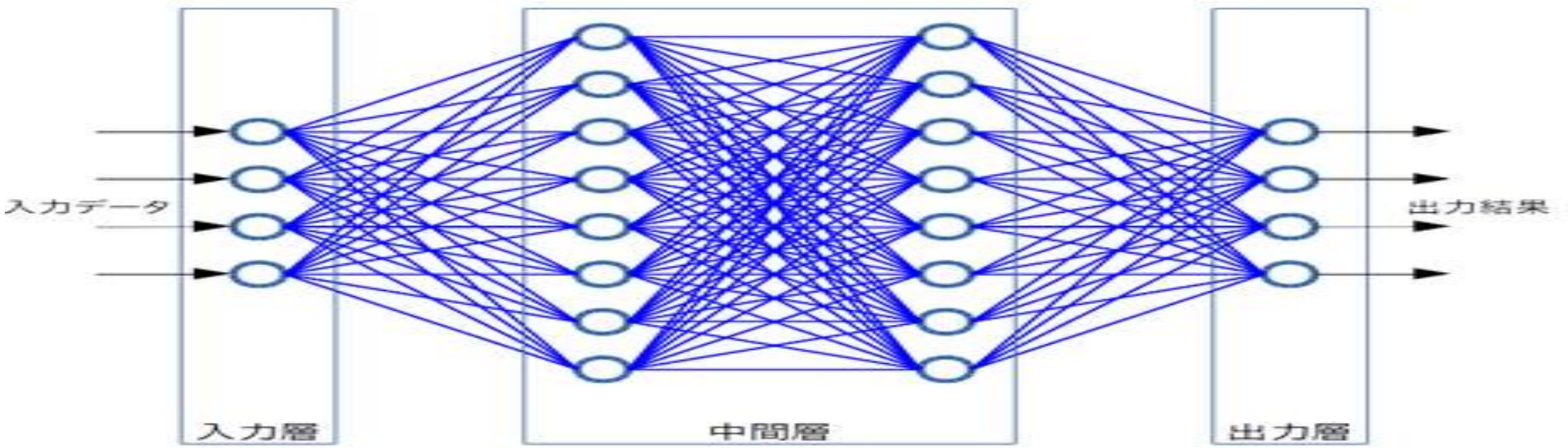
◆ネットワーク構造が極めて複雑なので、要因解析が困難

- 新たな研究や基本原理の解明が出来ない
- 理由がわからないと、結果の保証や適用限界が出来ない

◆ニクラス分類として用いると偶然相関の起きる確率が高い

この事実を避けるため、大量のサンプルの収集が大きな課題となっている

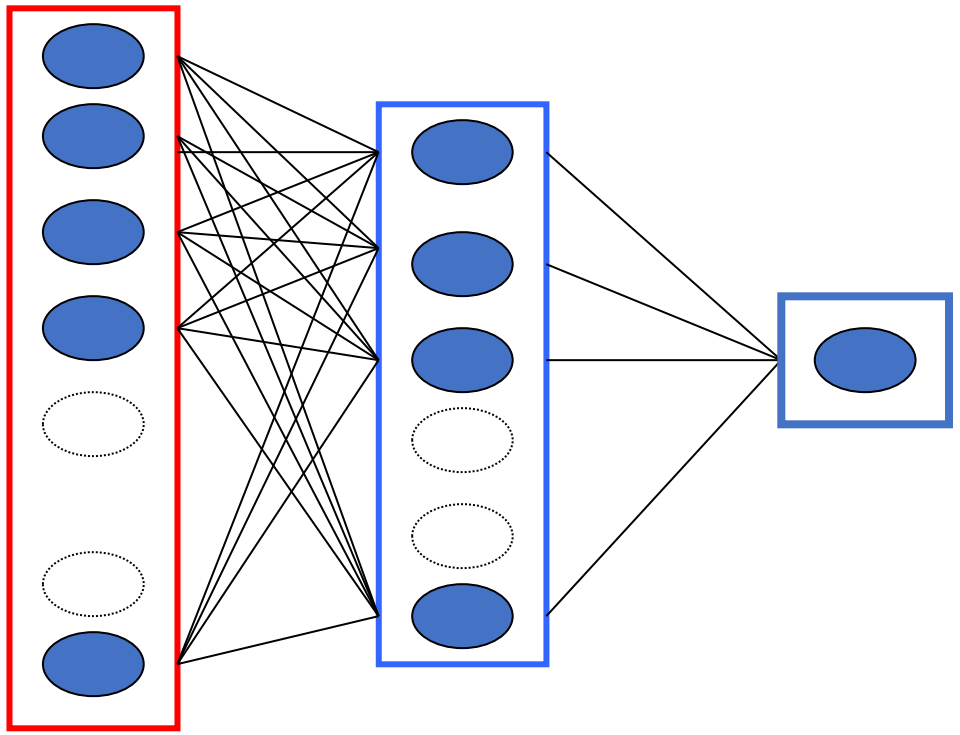
ＤＮＮ（Ｄｅｅｐ　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）ネットワーク構造

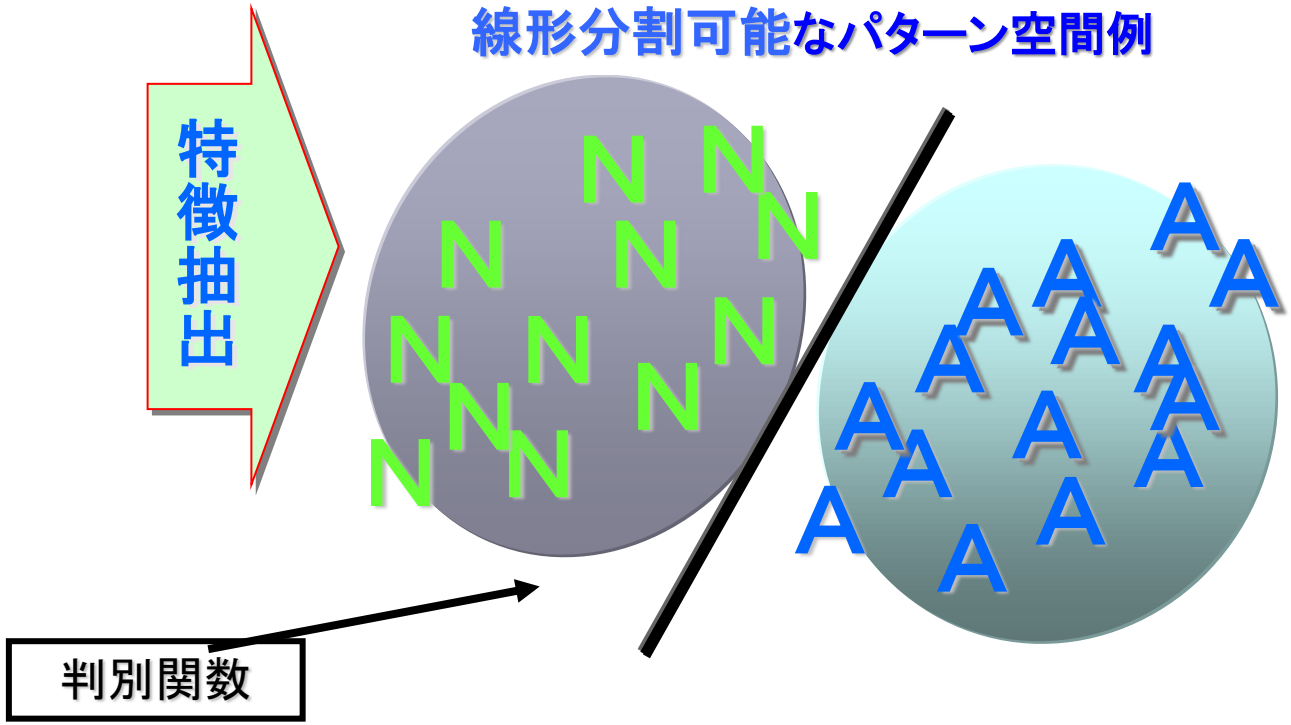


http://myinner.asia/archives/できるだけわかりやすく説明してみるという実験%EF%BC%9A深層学習%EF%BC%88ディープラーニング%EF%BC%89の基礎、ニューラルネットワーク

□機械学習における問題

ニクラス分類問題における過剰適合問題





空間にＡとＮを分ける情報を持たない

空間を形成するパラメータ群は
ＡとＮの識別根拠情報が見えにくい

分類目的を達成する空間構築による分類

空間にＡとＮを分ける情報を持つ

空間を形成するパラメータ群は
ＡとＮの識別根拠情報が見えやすい

□機械学習型人工知能

・機械学習型人工知能構築上での留意点

◇学習用に使うサンプル数が極めて大（化合物関連分野でのサンプル収集）

・データ解析手法として展開する場合、
サンプル数が少ないと偶然相関や過剰適合が起こり人工知能の信頼性が低下する

◇学習用サンプルでは、情報が偏らないようにする事が必要

・人工知能にかぎらず、データ解析という観点でもサンプルデータの偏りは危険

◇ネットワーク構造が極めて複雑なので要因解析ができない

・構造－活性/毒性/物性相関等の研究では、要因解析が極めて大事である

◇データクレンジング(Data Cleaning)が大事

・学習用サンプルデータは様々な形でのノイズがない状態であることが望ましい

□機械学習型人工知能

・機械学習型人工知能構築上での留意点

◇学習用に使うサンプル数が極めて大（化合物関連分野でのサンプル収集）

・データ解析手法として展開する場合、
サンプル数が少ないと偶然相関や過剰適合が起こり人工知能の信頼性が低下する

・ニューラルネットワークはパーセプトロン等と比較してネットワーク構造が複雑なため、
偶然相関や過剰適合が起こりやすい。
・深層学習はニューラルネットワークよりも更にネットワーク構造が複雑である。このため、
深層学習をデータ解析として利用する場合は、サンプル数を大きくすることが必須。

①世界一となったアルファ碁は、コンピューター同士での対局での強化学習を含めて、
全体で数千万局の学習をこなしている
②画像認識で飛躍的な認識率を上げた例では、数百万件の画像データ利用

□データサイエンス手法に関する制限事項や解析上での留意点

１．データサイエンス手法の適用には、データ解析という観点や解析信頼性を高く保つという観点で様々な留意点や制限事項を守る必要がある。

２．本日まとめた制限事項を意識しなくても、データサイエンスを実行すればそれなりの結果が出る。これがデータサイエンス実施の上での最も危険なことである。

３．分類率や予測率だけでデータサイエンスを実施すると、目先の指針だけが最適かされて解析信頼性は伴わない。

４．目先の指針に惑わされて、ソフトのハイパーパラメータ等を細かに変化させると、追加データでの解析を困難とし、解析の再現性を低下させる。

Importance of canonicalization : 一元一項対応

■ Extremely important concepts to keep in mind when dealing with compounds

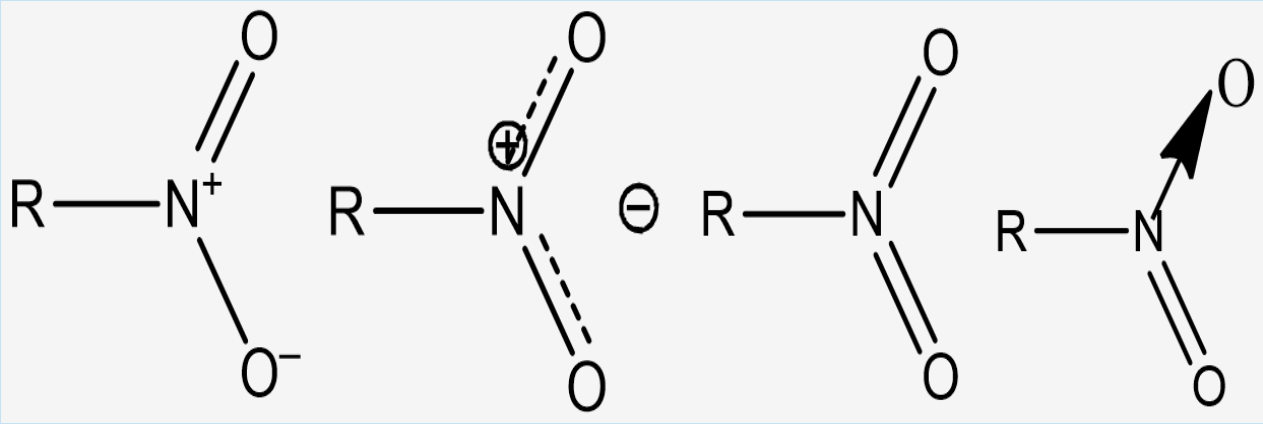
・ One-unit, one-item correspondence ⇒
There is only one notation (using the same notation) to specify one compound.
・ One-dimensional　multinomial support ⇒
There are multiple notations that specify one compound

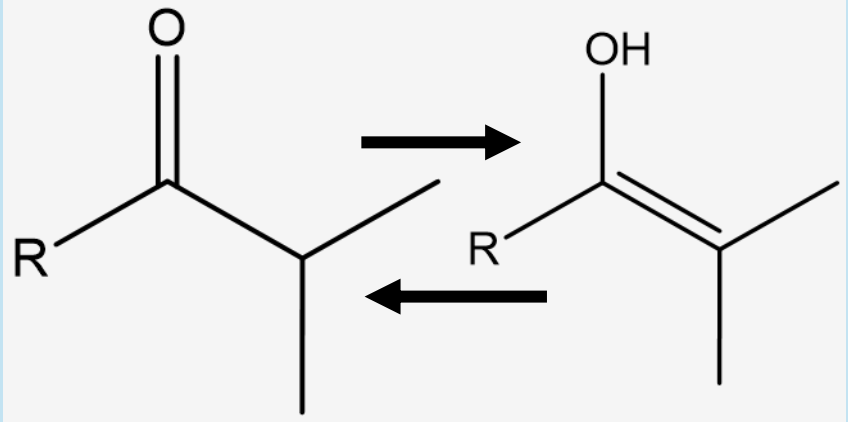
■化合物を扱う場合、常に留意すべき極めて重要な概念

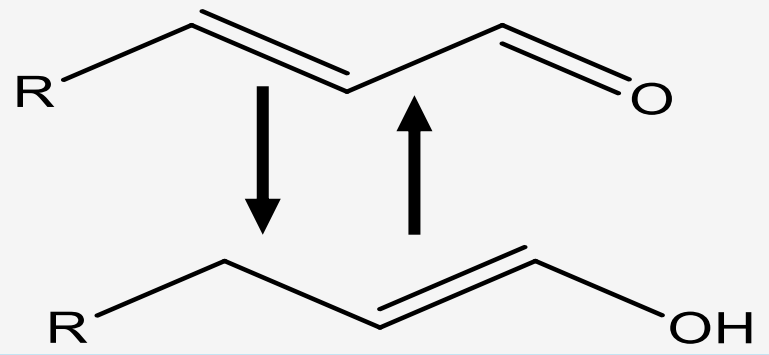
・　一元一項対応　⇒
一つの化合物を指定する表記は一つ（同一形式の表記法にて）しか無い
・　一元多項対応　⇒　一つの化合物を指定する表記は複数存在する

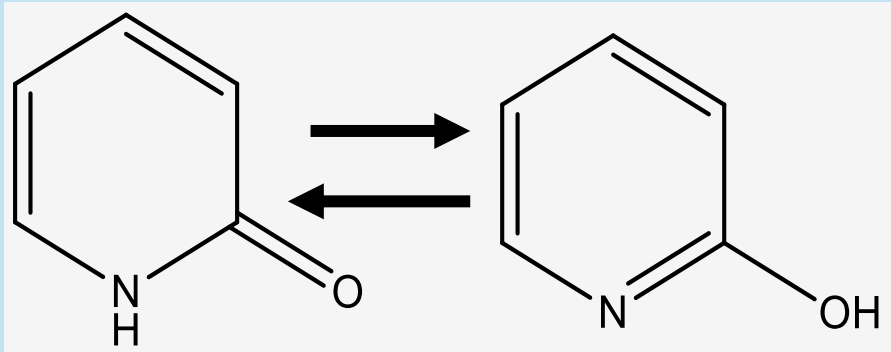
◇Compound-specific problems in the compound structure: Response to compound diversity is required

◇ Problem in compound structure: tautomer, nitro, aromatic



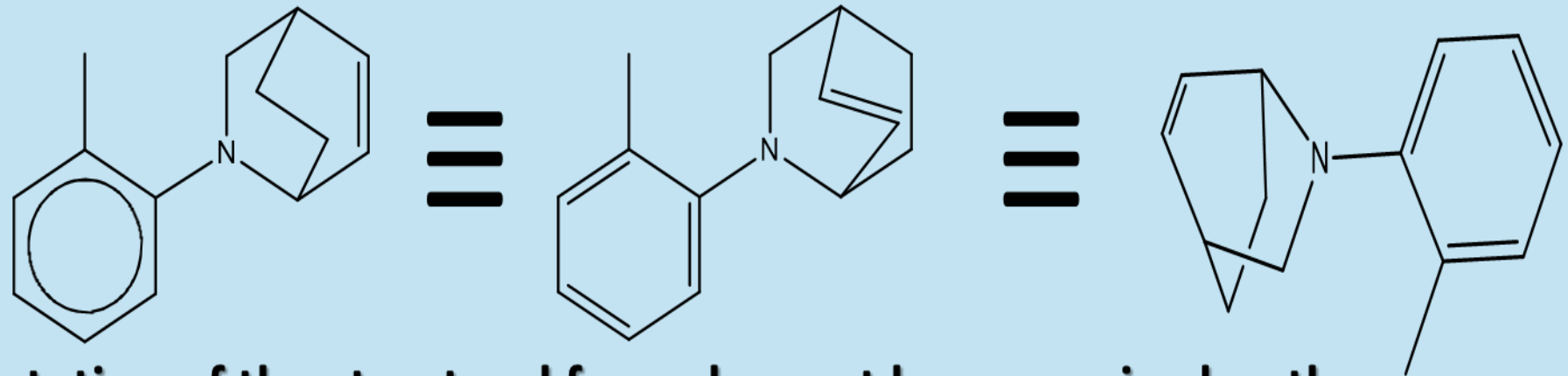




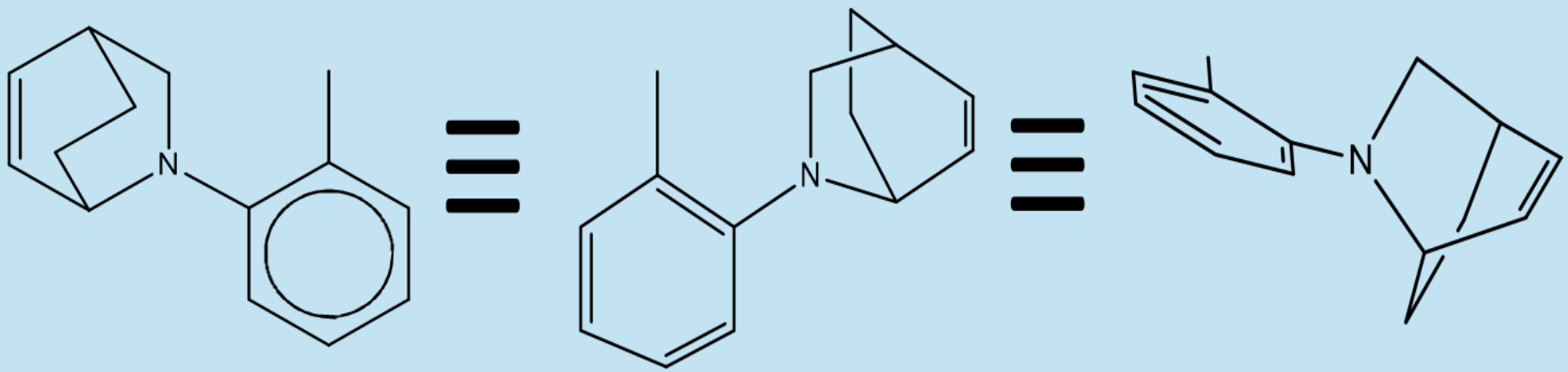


Many others

◇Necessity of canonicalization of compounds: Response to compound diversity



Any notation of the structural formula must be recognized as the same compound

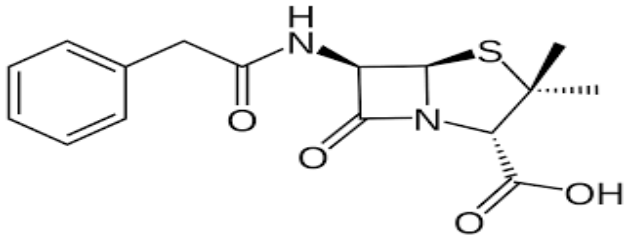


□化学分野の基本情報

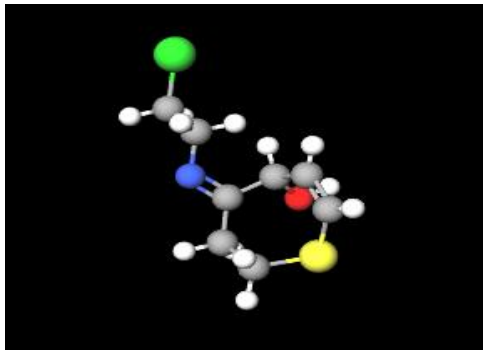
◇化学データ（アナログ）のデジタル化

＊化合物構造式（アナログデータ、イメージデータ、トポロジカルデータ）

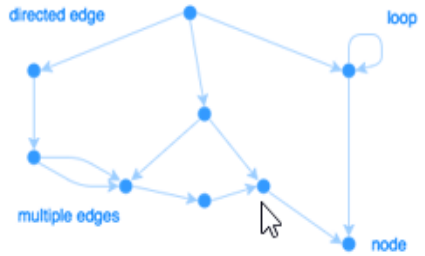
⇒　デジタル情報を基本とするコンピューターでは扱えない



二次元化合物構造表示



三次元化合物の
ボール＆スティック表示



Graph

Graph Convolutional Networks

0 0 1 0 0 1 1 0 1 1 0 1 1 0 0 0 1 0 1 0 0 0 1 1 0 1 0 1 0 1 1 1 1 0 0
1 0 1 1 0 1 0 0 1 0 1 1 1 0 1 0 0 1 0 0 1 1 1 1 1 1 0 0 1 1 0 1 0 0

□化学分野の基本情報

◇化学データ（アナログ）のデジタル化

＊化合物構造式（アナログデータ、イメージデータ、トポロジカルデータ）

⇒　デジタル情報を基本とするコンピューターでは扱えない

＊二次元及び三次元構造式

⇒　コンピューターは一次元で０/1のデータしか扱えない
（２/3次元は想定外）

＊コンピューターサイエンスによる検索及びデータ解析

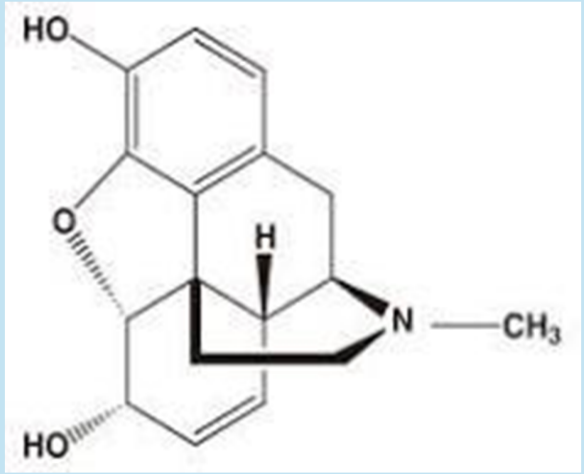
⇒デジタル情報を用いて展開されている

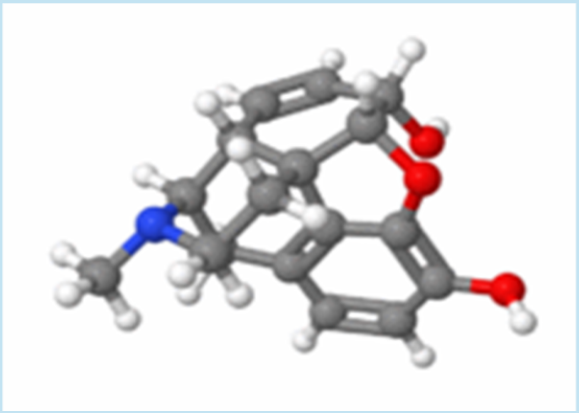
＊化学分野へのコンピューター適用の技術やサイエンスが存在

⇒**コンピューター化学（Computer Chemistry)**が展開されてきた

化合物操作上での問題：Problems related to compound manipulation / storage

◇ Diversity of compound notation: No unified information





■Chemical ID Number

CAS number:57-27-2

ATC code:N02AA01 (WHO)

PubChem:CID: 5288826

DrugBank:APRD00215

ChemSpider:4450907

KEGG:D08233

■compound properties

Chemical formula:C17H19NO3

■Reproducibility of chemical compounds: Linear notation of compounds

Compound name：Morphine

IUPAC: (5α,6α)-7,8-didehydro-4,5-epoxy-17-methylmorphinan-3,6-diol

SMILES: OC(C=CC1CC2N3C)=C(OC4C(O)C=5)C1C4(CC3)C2C5

InChIKey: InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1

□化学分野の基本情報

■化合物表記法の種類と違い（分子式、化合物名、WLN、Smiles）

線形による化合物表記

①CAS（Chemical Abstracts Service）番号 最新の登録化合物数：１億４千４百万化合物

- 番号は基本的に登録順で、左の数値、中央の数値を用いた通し番号がつけられる
- 構造や物性などとは関連付けることなく割り当てられ、番号に化学的な意味は持たせていない

異性体は異なる物質なので、CAS登録番号の割り当ても異なる。例えばD-グルコースは50-99-7、L-グルコースは921-60-8である。まれに、分子の種類全体に対して1つの CAS 登録番号が割り当てられることもある（全てのアルコール脱水素酵素は9031-72-5である）。

チェックディジットの計算式は次のとおりである。

CAS登録番号が $N_8N_7N_6N_5N_4N_3$-N_2N_1-R （R, N_i は各桁の0～9の数字、桁が存在しない場合は0とみなす）の場合、

$$R = (8 \times N_8 + 7 \times N_7 + 6 \times N_6 + 5 \times N_5 + 4 \times N_4 + 3 \times N_3 + 2 \times N_2 + N_1) \bmod 10$$

たとえば、水のCAS登録番号は 7732-18-5 なので、以下の通り5になる。

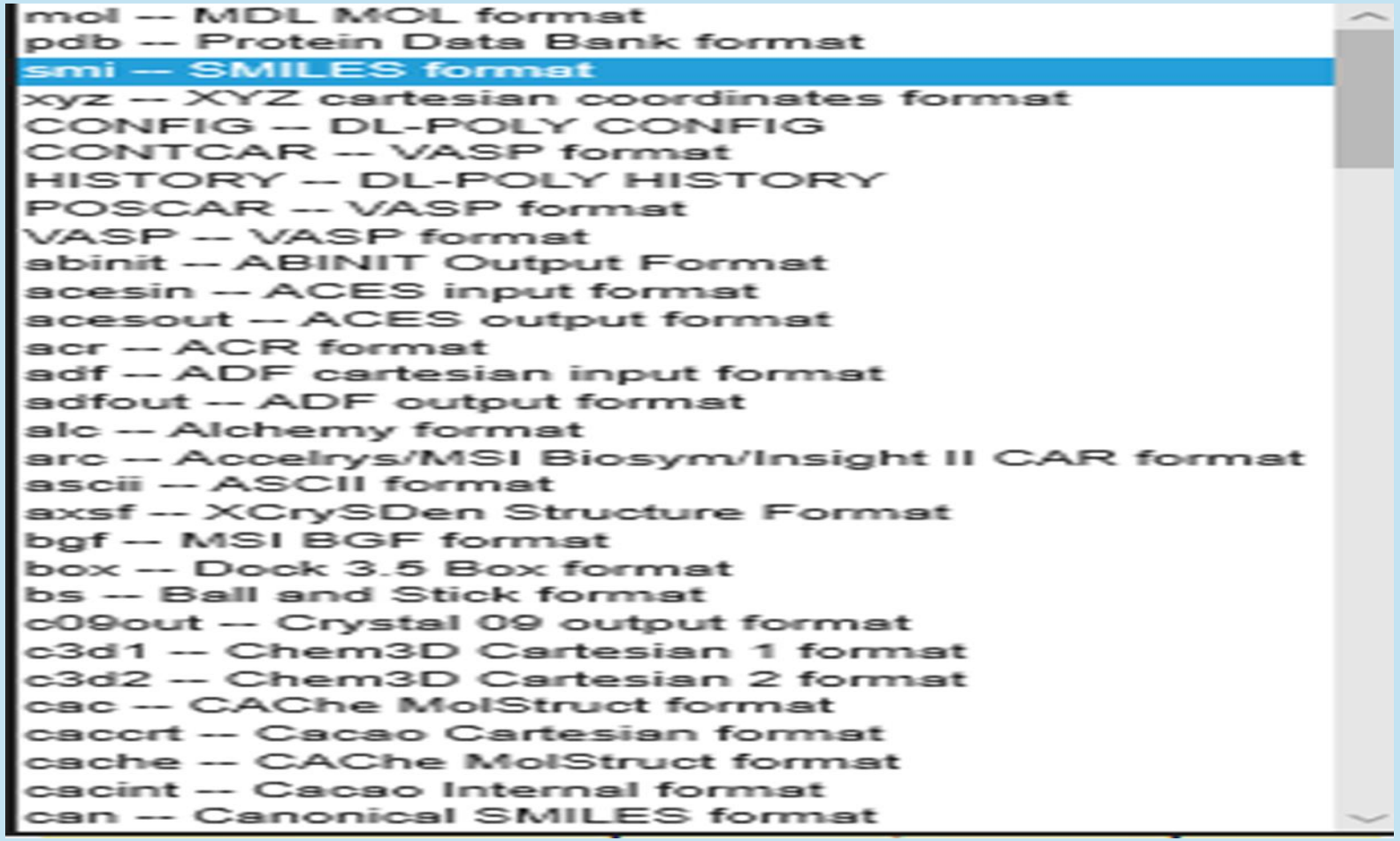
$$(6 \times \mathbf{7} + 5 \times \mathbf{7} + 4 \times \mathbf{3} + 3 \times \mathbf{2} + 2 \times \mathbf{1} + 1 \times \mathbf{8}) = 105$$

$$105 \bmod 10 = 5 \ (105 = 10 \times 10 + 5)$$

https://ja.wikipedia.org/wiki/CAS登録番号

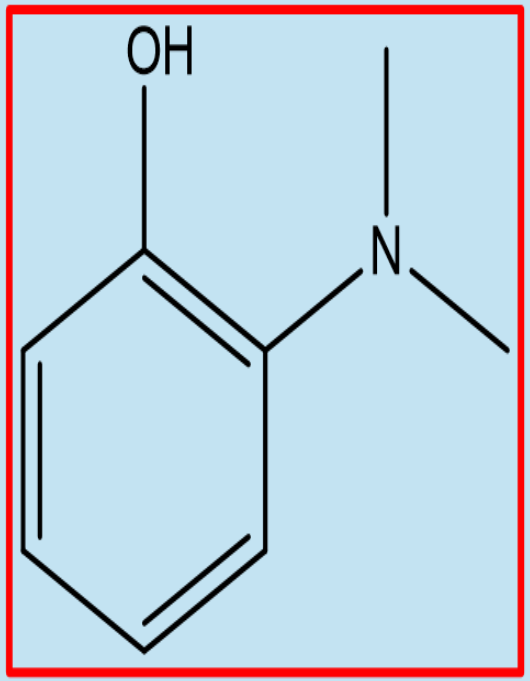
■Reproducibility of chemical compounds: Notation by connection table

List of file formats handled by the “OpenBabel system”



☐**Canonicalization is required to correctly perform compound searches**

There are many structural patterns in one compound. Compound does not hit in search.



SMILES 1: OC1=C(N(C)C)C=CC=C1 ;by **ChemDraw**

2:c1(O)c(N(C)C)cccc1 ;by **Ecosar**

3:C1=CC(=C(C=C1)N(C)C)O ;by **QSAR Toolbox**

4:CN(C)c1ccccc1O ;by **OpenBabel**

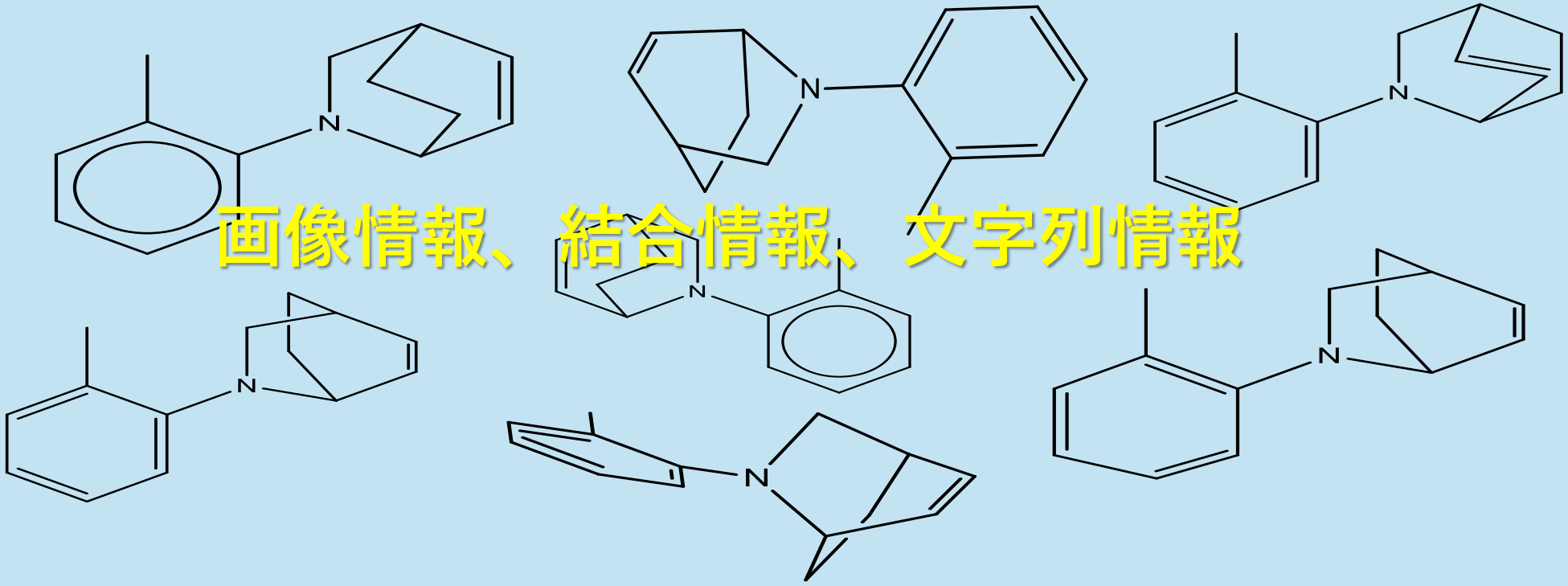
5:C1=CC(O)=C(N(C)C)C=C1 ;Manual Input by Yuta

6:C1(O)=C(N(C)C)C=CC=C1 ;Manual Input by Yuta

□二次元化合物構造式の変化性問題

◆全く同じ化合物が作画状態の違いで異なる図となる
1．化合物の**方向性**の違い（上下／左右／表裏）
2．**表記**の違い（芳香族結合、ブリッジ構造、他）

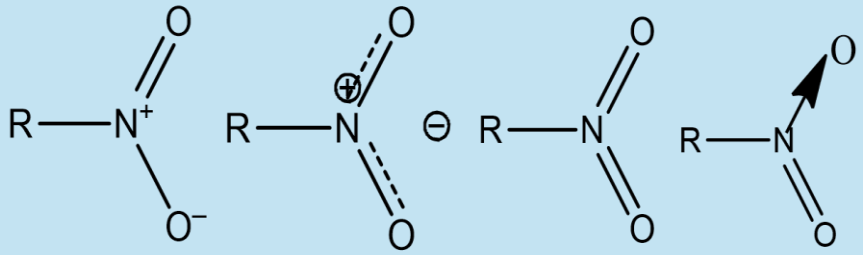
◆コンピューター上では全く同じ化合物と認識されるか？

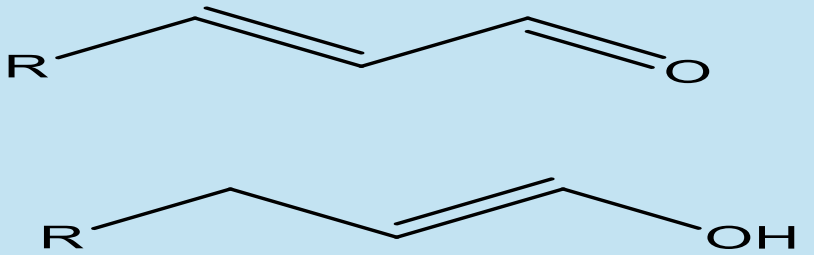


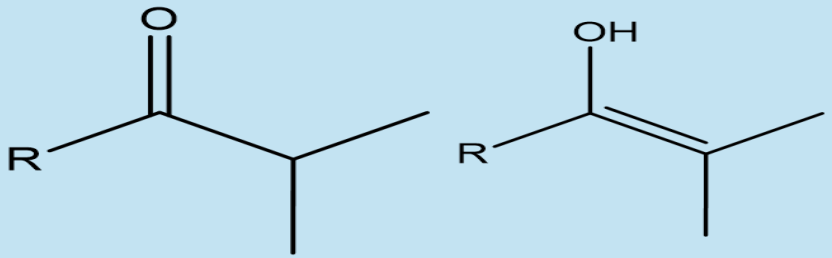
□化合物構造表記の多様性に関する問題

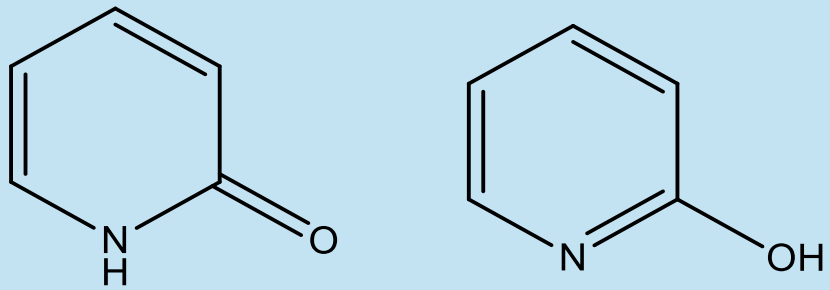
◇ Problem in compound structure:

・ tautomer　・nitro　・aromatic　・salt









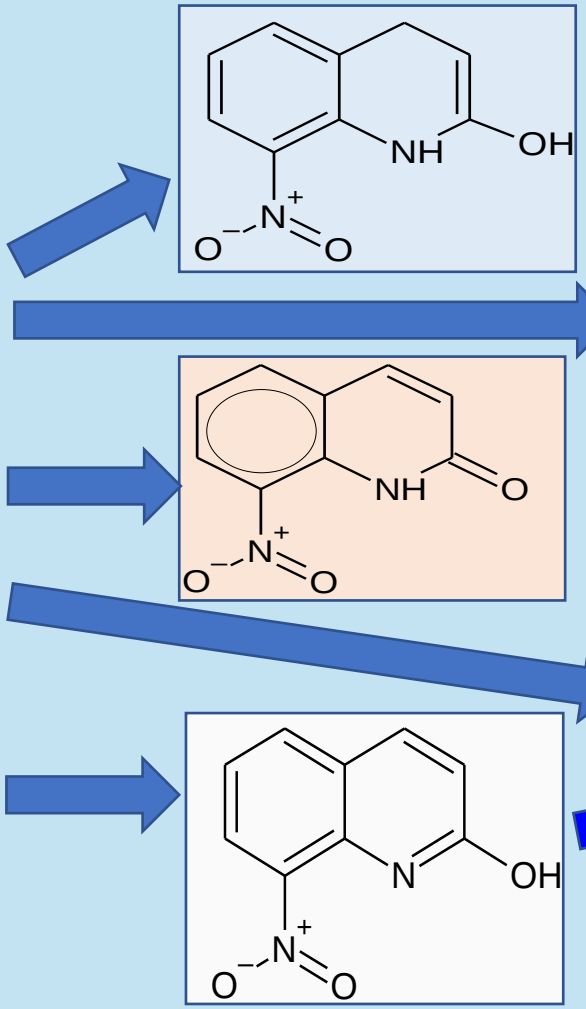
・作画者の考えや習慣、用途等により変化する
・表記を間違えたわけではなく、すべて正しい

一元多項対応(One to Many relation)

In Silico Data
Miracles by the KY-methods

Original Compound

起源化合物



No reliability & Accuracy

Data Science & Artificial intelligence

□化合物のシステム開発や利用上での化合物操作上での問題点

◆以下の内容に関して、システム利用目的に応じて対応必要

1．化合物表記手法の変化性
- 多種多様の表記法が存在
- 同じ表記法であっても内容が異なる

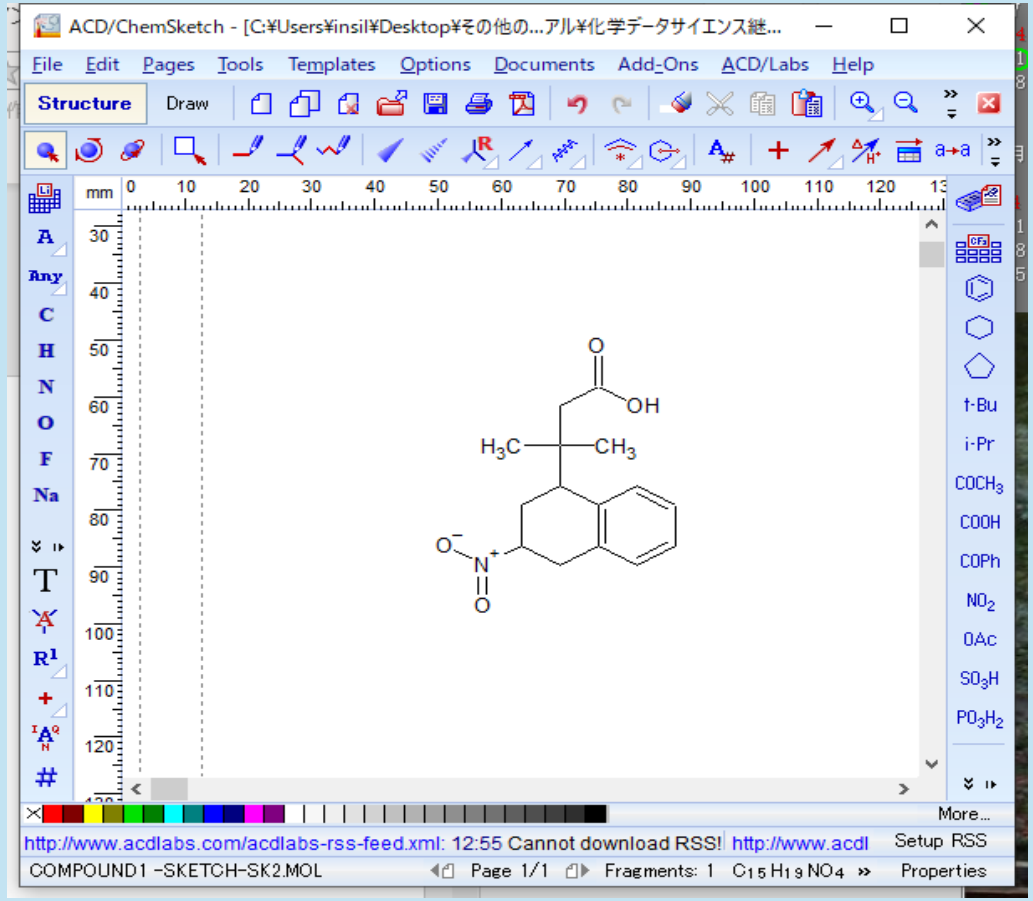
2．化合物入力時の変化性
- 二次元構造式の変化性
- 三次元構造式の変化性（ローカル／グローバル）

3．化合物構造式の多様性
- 同じ化合物に正しい表記が複数存在

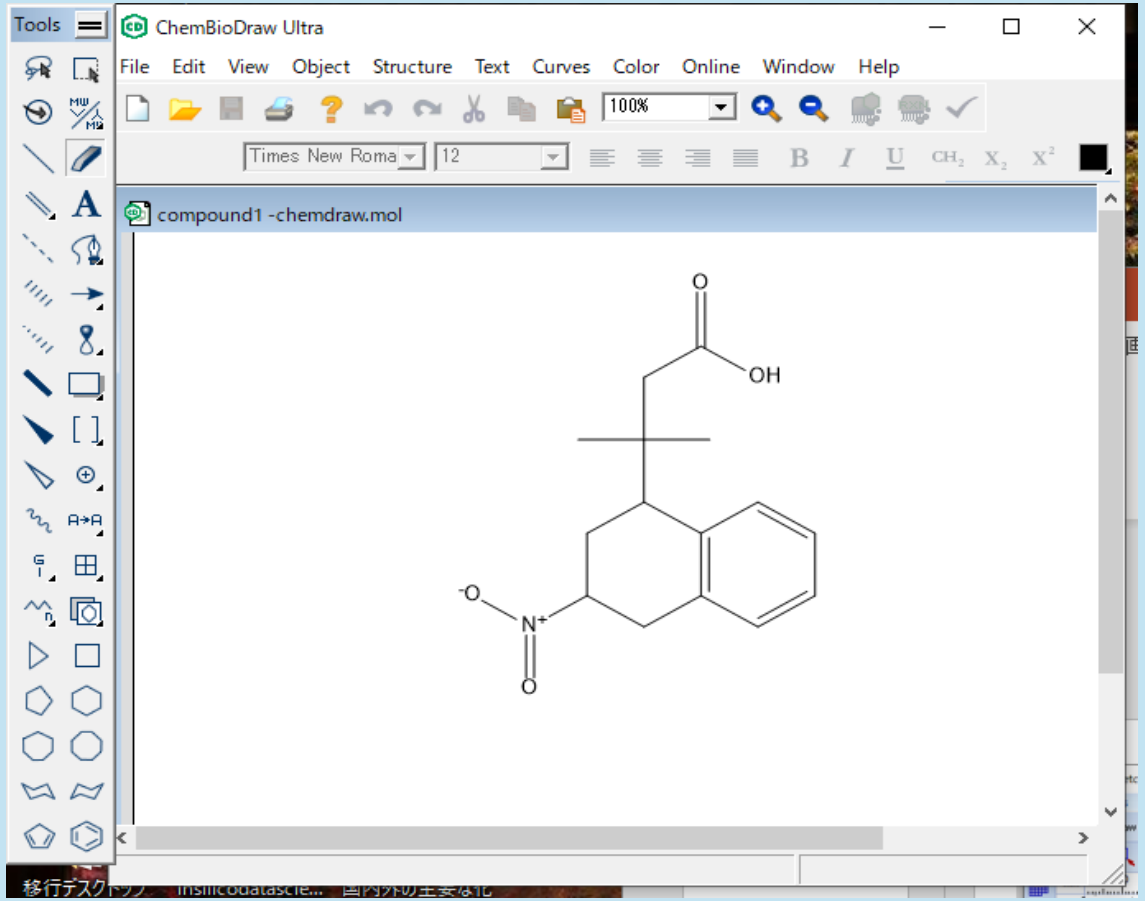
□二次元化合物構造式の作画（異なるソフト）

In Silico Data
Miracles by the KY-methods

ACD/ChemSketch



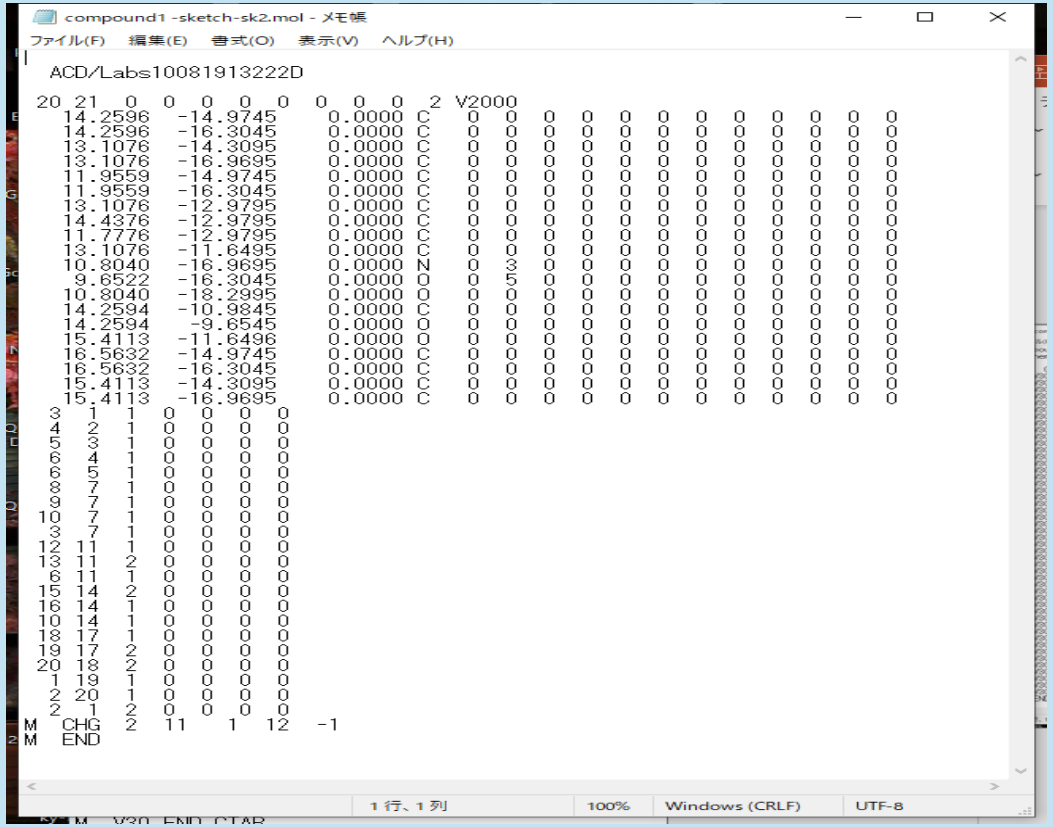
ChemBioDraw



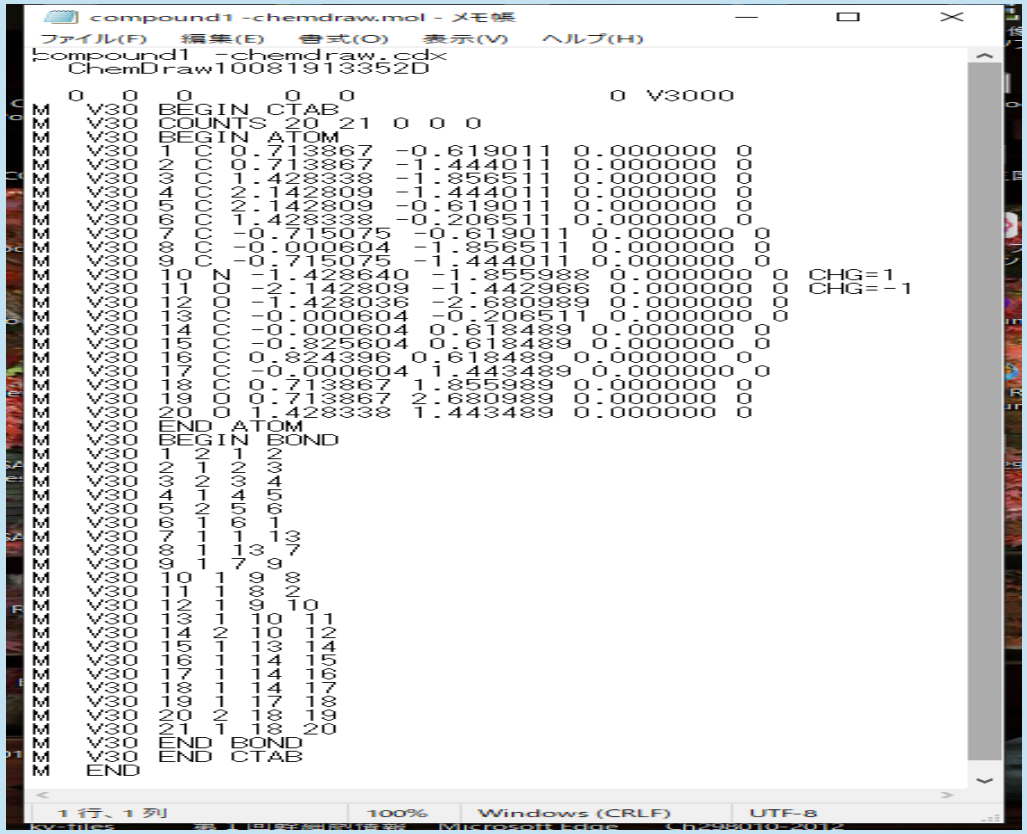
□二次元化合物構造式の同一Molファイル

◆同一ファイル形式でも、互換性に留意が必要
・システム依存性、　・バージョンの変化、記述桁数の違い、他

ACD/ChemSketch

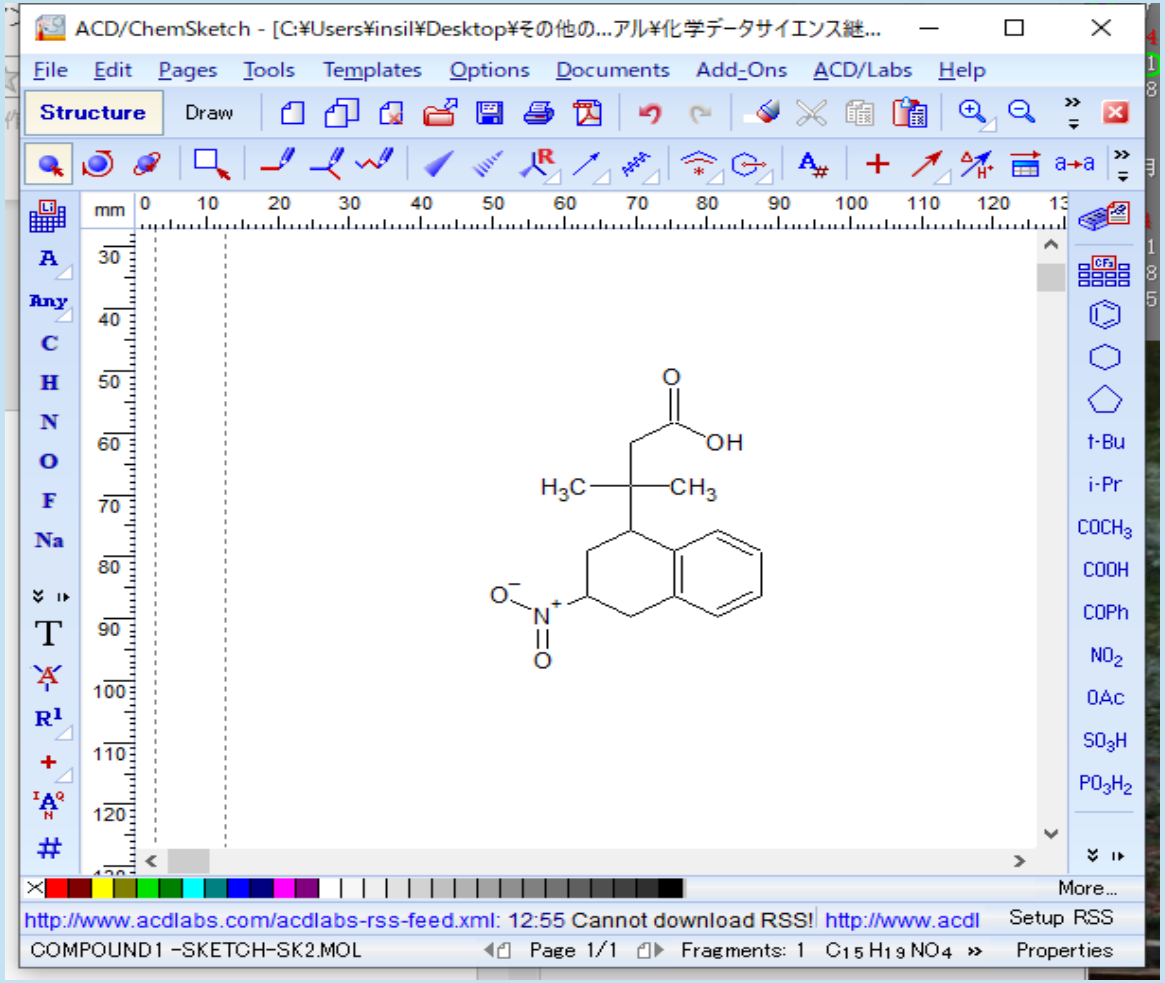


ChemBioDraw

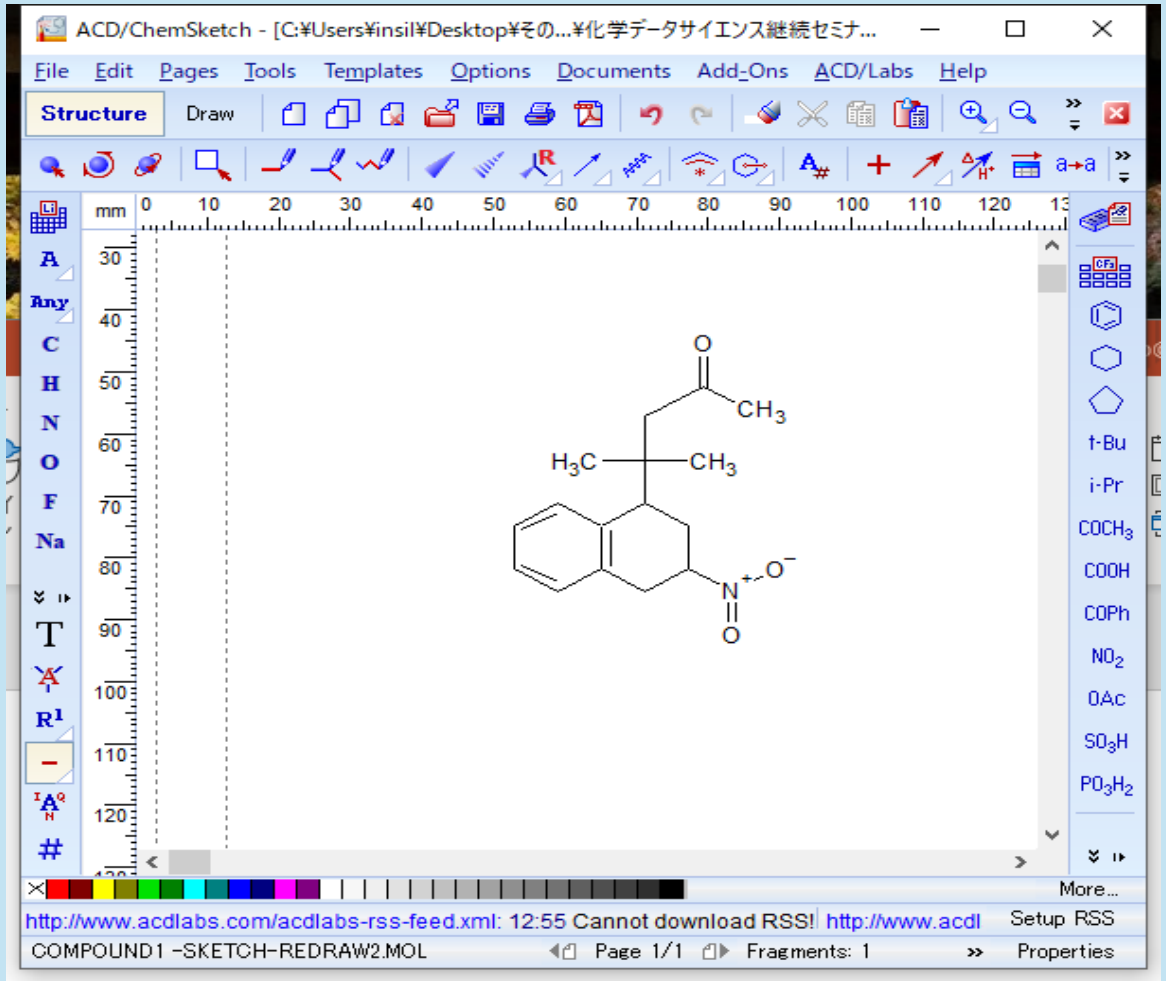


□二次元化合物構造式の作画（同一のソフト）

ACD/ChemSketch

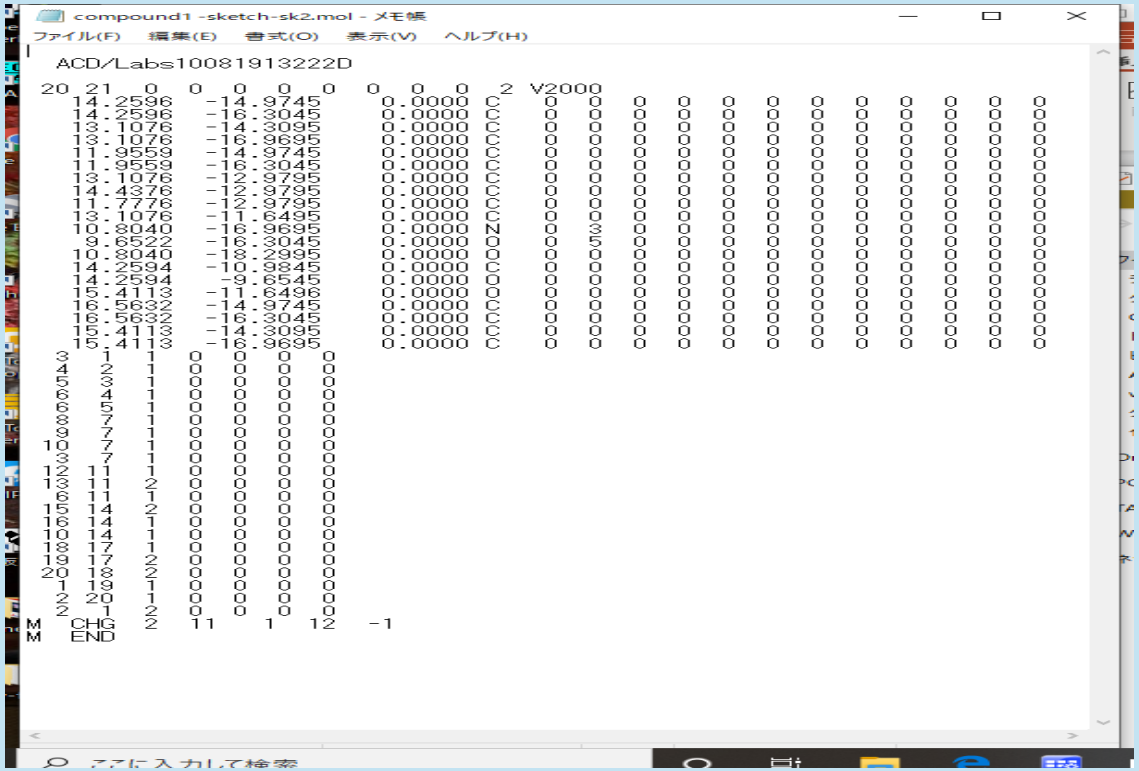


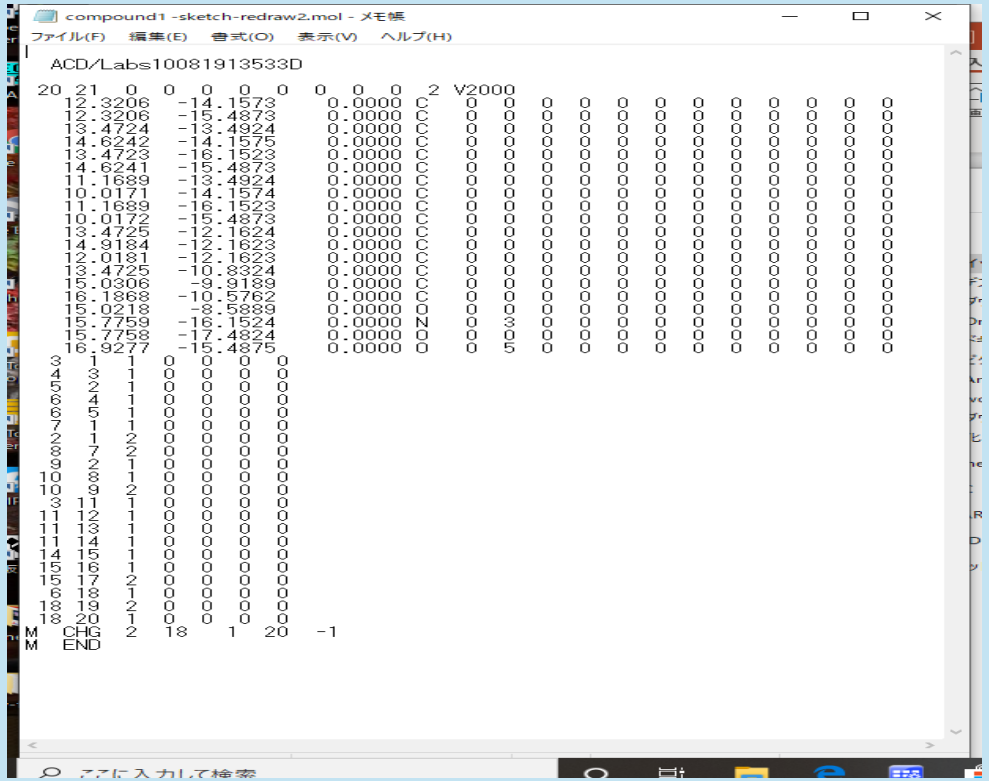
ACD/ChemSketch



□二次元化合物構造式の同一Molファイル

◆原子の番号付けが異なる
同一システム作成の同一化合物であっても、**原子番号が異なる**
原子番号の一元化が必要（Morganアルゴリズム、SEMA等）
◆化合物の種々検索での問題解決が必要





□化合物を扱う時のデータサイエンス手法に関する制限事項や解析上での留意点

In Silico Data
Miracles by the KY-methods

1．コンピュータ上で化合物を扱う時は、「一元一項」対応が最も重要な考えである

2．一つの化合物がコンピュータ内で、違った表記で登録される可能性を常に考える

3．たとえ構造式が異なっても、化学的には全くの同一化合物であることはよくあるが、
コンピュータ技術上では異なった化合物として扱われることが多い。注意が必要である。

4．同一化合物が異なった形式で登録されると、データ解析は矛盾をきたし、間違った結論に至る

5．化合物表記は、その利用目的や時代の要請に応じて多種多様な形式をとる。利用目的の異なる分野の
化合物データを取り扱う時には、内部情報の違いや扱いに注意が必要

6．構造式作画ソフトは便利であるが、個々の特徴や、作画する人の書き方の違い等吸収できるかチェック必要

7．化合物の立体の扱いは重要で、立体情報をデータ解析でどのように扱うかも注意が必要な事項である

8．ビッグデータとして化合物DBを複数融合する時は、化合物の扱いの差に注意が必要である

TS-02：
まとめと「オートノマス創薬」提案に続く