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

1. 人工知能概論と化学分野の人工知能

・時代の変化と人工知能の歴史

生命科学分野での人工知能への期待分野と適用／成果

	化学／創薬	バイオ	医療
適用分野			
人工知能	初期⇒ ルールベース型 現在⇒ 機械学習型（深層学習）		
期待成果	新薬デザイン、最適化、ドラグ・リポジショニング、毒性評価、その他	遺伝子解析、遺伝子探索、発現プロフィール解析、SNPs探索、その他	自動診断、画像解析、音声解析、その他

1. 人工知能概論と化学分野の人工知能
・時代の変化と人工知能の歴史

In Silico Data
Miracles by the KY-methods

□最新／現役／話題の人工知能システム

☆ルールベース型人工知能

DEREK　⇒　化合物の毒性評価システム

☆コグニティブコンピュータ

Watson (IBM) ⇒クイズ番組で勝利
医療関連分野で実績を出しつつある

☆機械学習（深層学習）型システム

アルファ碁（グーグル）⇒
世界のトッププロ棋士（イ・セドル）に４勝一敗で勝利

- 学習回数：数千万局＞三千万局（自己対局）
- ルールの自動獲得：囲碁のルールを自動的に学習した？

1. 人工知能概論と化学分野の人工知能

・時代の変化と人工知能の歴史

In Silico Data
Miracles by the KY-methods

☆機械学習（深層学習）型システム

アルファ碁（グーグル）⇒

世界のトッププロ棋士（イ・セドル）に４勝一敗で勝利

勝因？：**ルール型対局**から一種の**パターン認識型対局**に変えた

この結果、対局勝利まで今後十年かかるといわれた評価を覆した

口対局の条件と特徴的結果

- 学習サンプル数：数千万局＞三千万局（自己対局）
- ルールの自動獲得：囲碁のルールを自動的に学習した？

上記事実から受ける期待イメージ：

大量のデータを用いて深層学習させると、新規で何らかの重大なルールが発見でき、新しい研究に繋がるのではないか？

□　化学分野での人工知能の歴史

当初から現在まで

ルールベース型人工知能

最初に人工知能として開発。
実用システムや研究システムが
多数開発済み。
技術的にほぼ完成状態。
長所や欠点が見えている。
研究的に新規性が殆ど無いため、
論文になりにくい状況

現在注目中
機械学習型
（ニューラルネットワーク）
人工知能

機械学習型人工知能

システムとしての実績はないが、
今後の展開が期待されている
⇓
深層学習
未知要素が多く、期待値が高
論文になりやすい状況

□　機械学習型人工知能の全体的歴史と主要トピックス

□人工知能に注目させたトピックス：成功事例：アルファ碁
□人工知能に注目させたトピックス：失敗事例：チャットボット「Tay」
□人工知能の歴史：多変量解析／パターン認識と人工知能

◆過古の多変量解析/パターン認識と人工知能との関係
殆ど関連性が無く、全く別の研究分野として扱われてきた
唯一、単層型のパーセプトロンが、分類機として使われていた

◆現在における多変量解析/パターン認識と人工知能との関係
現在の深層学習を基本とする人工知能は、ニューラルネットワークが基本。
殆どの場合、ニューラルネットワークは多変量解析／パターン認識に分類される。この結果、両分野の境界は殆ど存在しなくなった。

人工知能に注目させたトピックス

□ 成功事例

AlphaGo（アルファ碁）が人間に打ち勝って世界一になった。
人間がコンピュータに勝てる最後の分野の神話が崩れた

碁の学習アルゴリズムに深層学習を適用していた

留意点：学習に使われた対局数が数千万という数に達している
サンプル数が少ない場合は成果を期待しにくい

人工知能の成功トピックスに隠れた事実

○ 極めて多数のサンプルが必要

- ネットワーク構造が複雑なため、過剰適合の回避に極めて多数のサンプルが必要
- 学習数が少ないと強くならない

○ 要因解析実施が極めて困難

- ネットワーク構造が複雑なため
- なぜアルファ碁が強いのかがわからない

1. 人工知能概論と化学分野の人工知能

・時代の変化と人工知能の歴史

In Silico Data
Miracles by the KY-methods

■ **失敗事例**

人工知能の対話型ロボット「Tay」が**ヘイト発言**を乱発

↓

Twitter等の内容から**ヘイト発言を学習**してしまった

留意点：人工知能における学習サンプルの重要性
学習内容により人工知能は大きく変化する

人工知能の失敗事例に学ぶべき事実

人工知能は学習内容に忠実に行動する（善人にも悪人にもなる）

○ 学習用サンプルの品質や内容の吟味

・望ましい判断が出来るように学習させる
ノイズ情報を極力避けて学習

○偏りのない学習をさせる

・学習には、正解と誤答の両方が必要

例：文献データは成功事例。
文献データを多数集めても、失敗事例データが無いと、人工知能は正しい評価ができない

人工知能と多変量解析／パターン認識

現在の機械学習型人工知能は多変量解析／パターン認識と基本は同じ

○ 多数のサンプルが必要

・過剰適合や偶然相関の回避

○学習には、正解と誤答の両方必要

・人工知能も多変量解析／パターン認識も
機械学習により知識や情報を得る
・学習内容が偏っていると、結果も偏ってくる
・多変量解析／パターン認識ではサンプル数の
多いクラスに振り分けられる

1. 人工知能概論と化学分野の人工知能

・人工知能の二大手法（ルールベース型と機械学習型）

化合物構造式中心で展開される化学と、
数値／文字中心で展開される自人工知能の
ギャップの解消

□ルールベース型人工知能：過去から現在

化合物の情報学的操作に関する技術は
LHASAシステムの開発過程でほとんど解決済

□機械学習型人工知能：現在

機械学習で化合物を扱う技術は、化学多変量解析／
パターン認識（ケモメトリクス分野）でほとんど解決済

構造式を細かに扱わない深層学習に問題あり

1. 人工知能概論と化学分野の人工知能

・人工知能の二大手法（ルールベース型と機械学習型）

In Silico Data
Miracles by the KY-methods

人工知能言語

1958:LISP(List Processor)
1972:Prolog
1994:Python

ルールベース型人工知能

＊第五世代コンピュータ（日本）

従来型人工知能
実用システム多数

多変量解析/パターン認識

重回帰、パーセプトロン、PCA、
クラスタリング、他

機械学習発展・新アプローチ

ニューラルネットワーク、
遺伝的アルゴリズム、ファジイ、他

深層学習開発／展開

新世代
人工知能

二種類の人工知能

■知識整理および適用型

ルールベース型人工知能

解決すべき問題点：
・目的解決に適したルール作成
・ルール間の階層、衝突回避
・エキスパートの存在必要

□発見型および要因解析型

機械学習型人工知能
ニューラルネットワーク
深層学習

解決すべき問題点：
・データ解析上の問題点
過剰適合、偶然相関、クラス分布、欠損データ、他
・解析手法の特性／限界
・解析結果の解釈

1. 人工知能概論と化学分野の人工知能

・人工知能の二大手法（ルールベース型と機械学習型）

人工知能の変化
時代の変化による

◇ルールベース型人工知能：

人工知能の初期に実施された。
限界；適用可能な妥当なルールベースを作りこむことが難しい

IoT等の発展により、ビッグデータ時代に突入

◇機械学習型人工知能：最近のメインアプローチ

大量のデータを用いて学習し、自動的に知識表現を獲得
ニューラルネットなどの機械学習手法が発達
最近のディープラーニング等が展開されている

1. 人工知能概論と化学分野の人工知能

・人工知能の二大手法（ルールベース型と機械学習型）

In Silico Data
Miracles by the KY-methods

化学分野で現在展開されている人工知能システム

□歴史的に化学関連分野への人工知能適用の歴史は長い
　化学分野では数式に乗らない事項が多く、経験則が重要となることが多い
　⇒人工知能が活躍する地盤がある

□適用事例は多い
・機器スペクトルデータの解析支援システム
・有機合成支援システム
・毒性予測システム
・構造−活性相関支援システム
・創薬化学者支援システム
・その他
　従来より展開されてきた化学分野の人工知能システムは、その展開上化学的なノウハウや考え方等のアナログ的な内容を、デジタルに変換する事が必要
　⇒ルールベース型

人工知能による毒性予測関連支援システム：ルールベース型人工知能

DEREK: **D**eductive **E**stimation of **R**isk from **E**xisting **K**nowledge

HazardExpert:

RIPT: **R**ule **I**nduction for **P**redictive **T**oxicology

TOX-MATCH:

DART: **D**ecision **A**nalysis by **R**anking **T**echniques

1. 人工知能概論と化学分野の人工知能

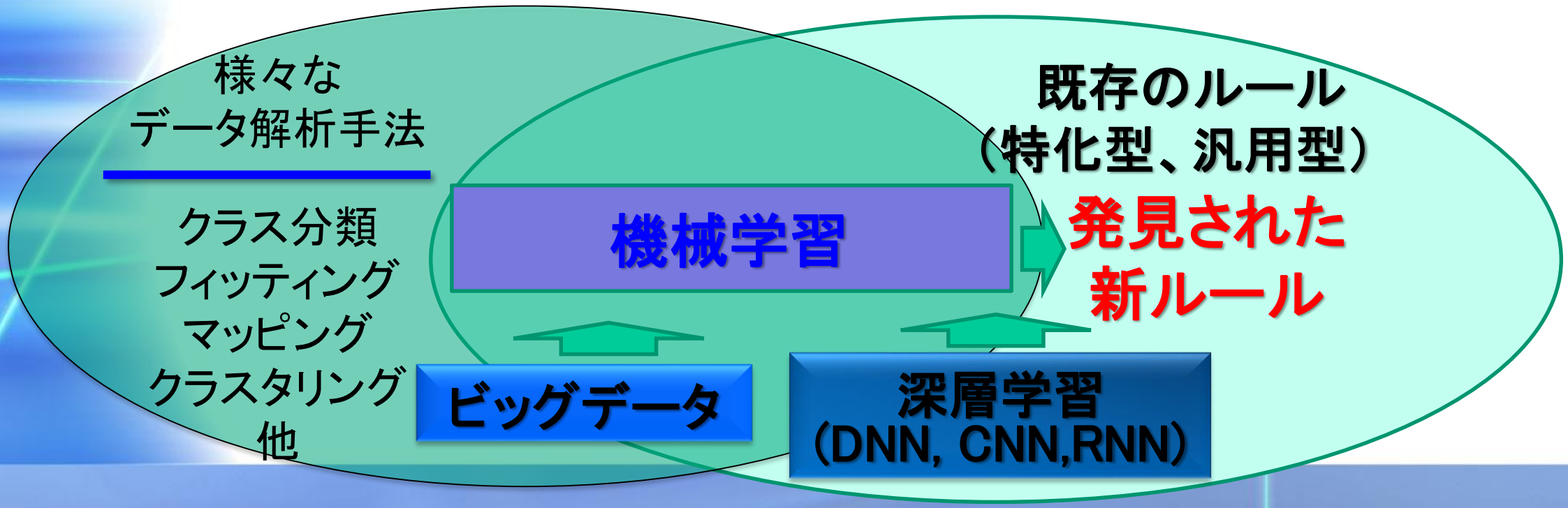
・ケモメトリックスと人工知能の関係

In Silico Data
Miracles by the KY-methods

現在における多変量解析/パターン認識と人工知能との関係

多変量解析/パターン認識と人工知能は
機械学習により繋がっている

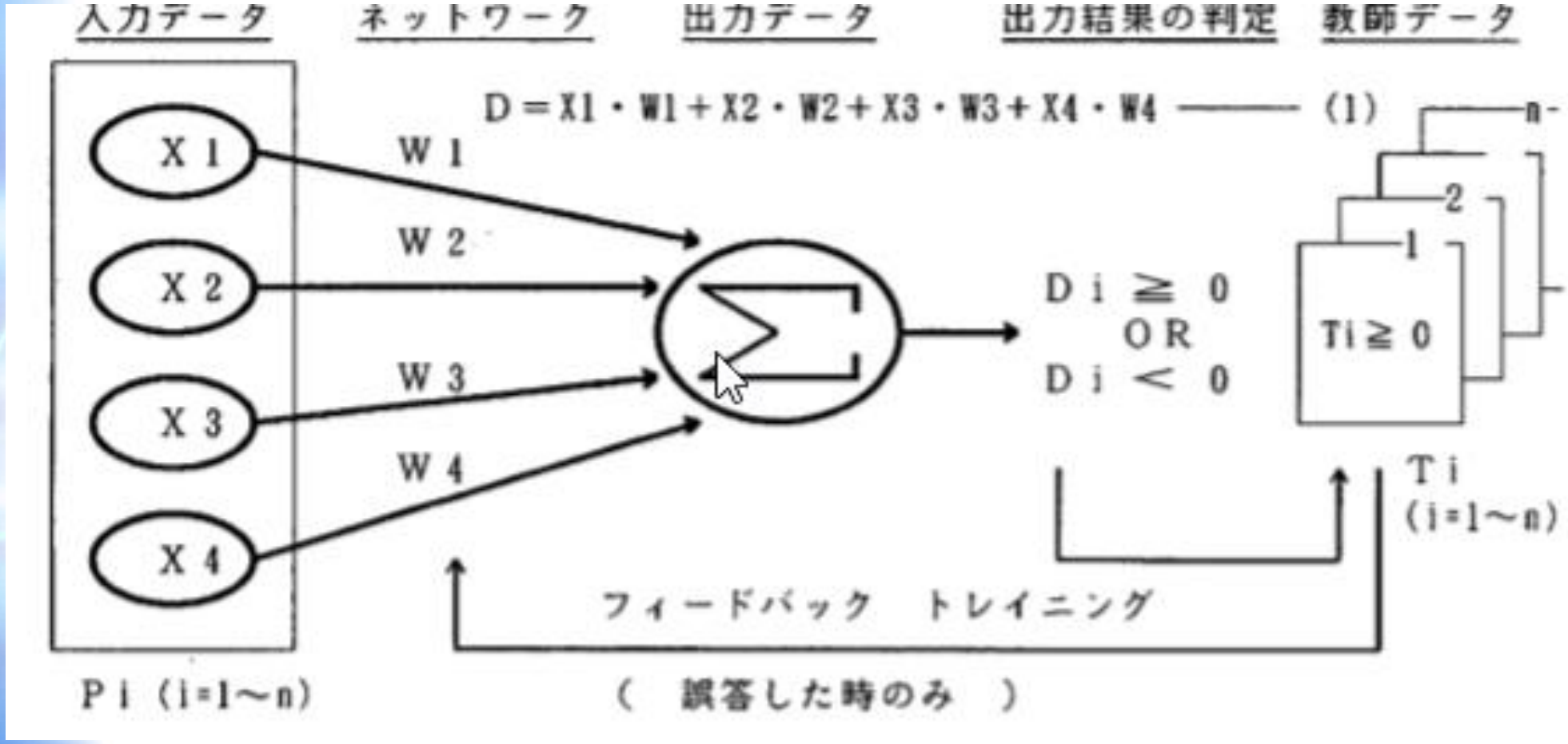
多変量解析/パターン認識　　**人工知能**



1. 人工知能概論と化学分野の人工知能

・ケモメトリックスと人工知能（ニューラルネットワーク型）の関係

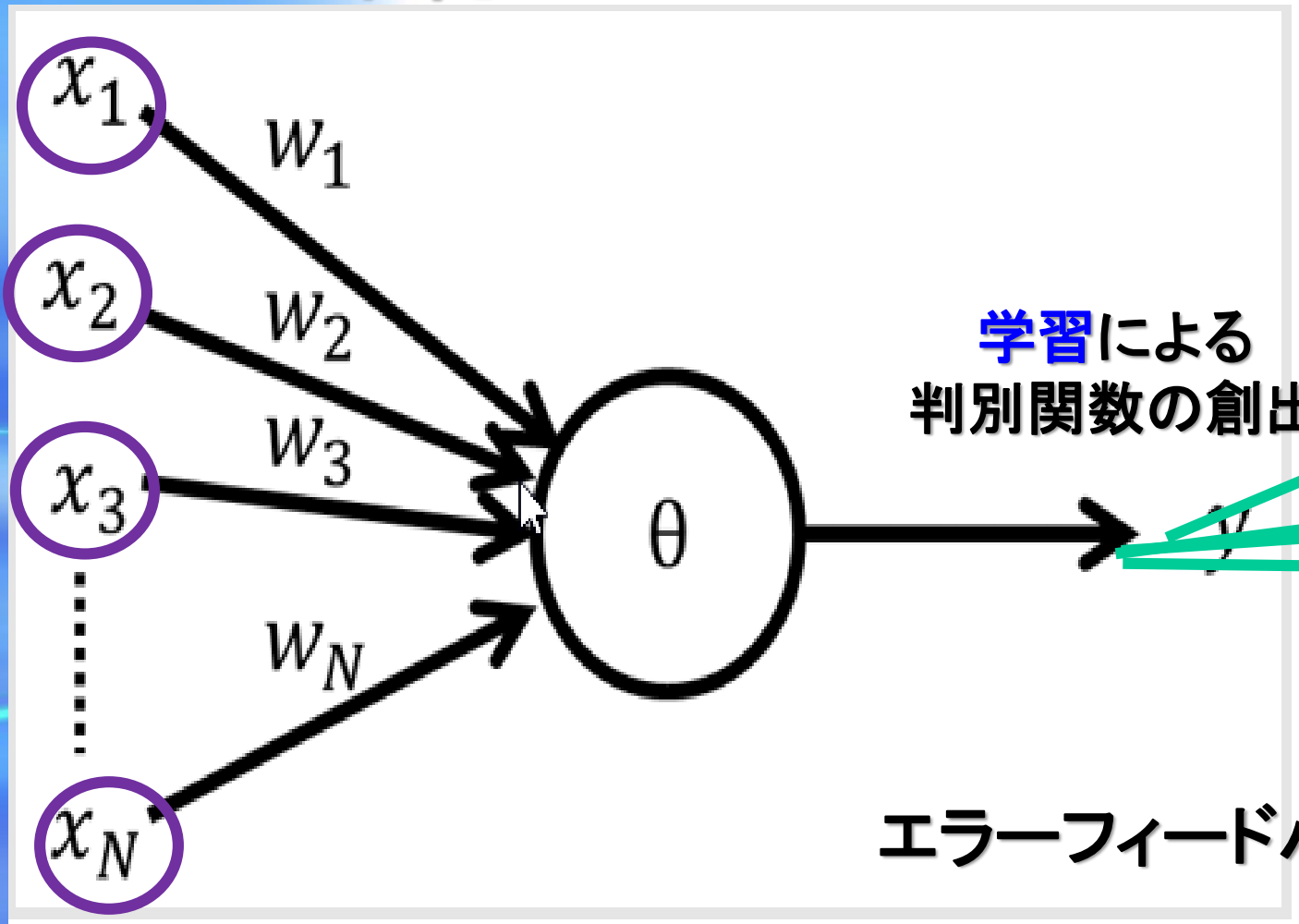
パーセプトロンの学習アルゴリズム



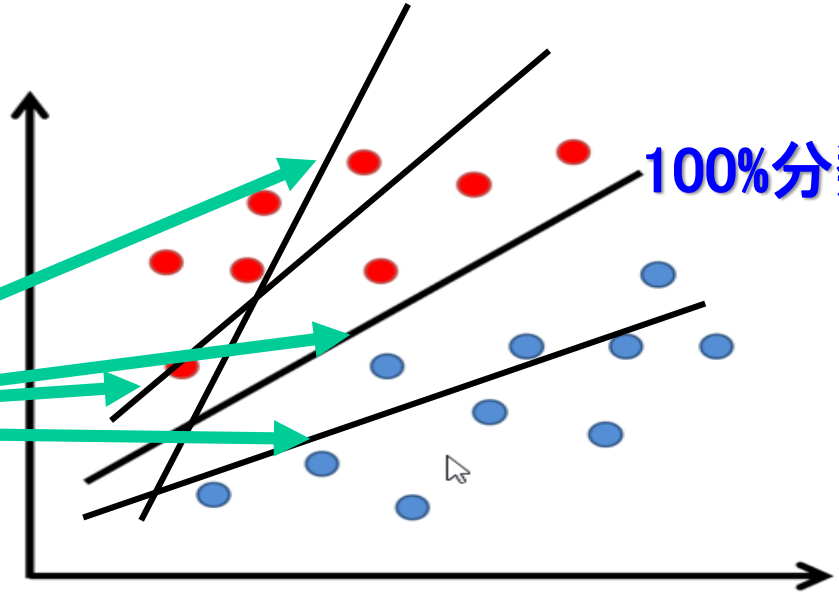
1. 人工知能概論と化学分野の人工知能

・ケモメトリックスと人工知能の関係

単純パーセプトロン



学習による
判別関数の創出



エラーフィードバックトレーニング

1. 人工知能概論と化学分野の人工知能

・ケモメトリックスと人工知能の関係

エラーフィードバックトレーニングのイメージと計算式

判別関数の修正

$$D' = -D \quad (3)$$

$$Wn \cdot Xp = -Wo \cdot Xp \quad (4)$$

$$Wn = Wo + C \cdot Xp \quad (5)$$

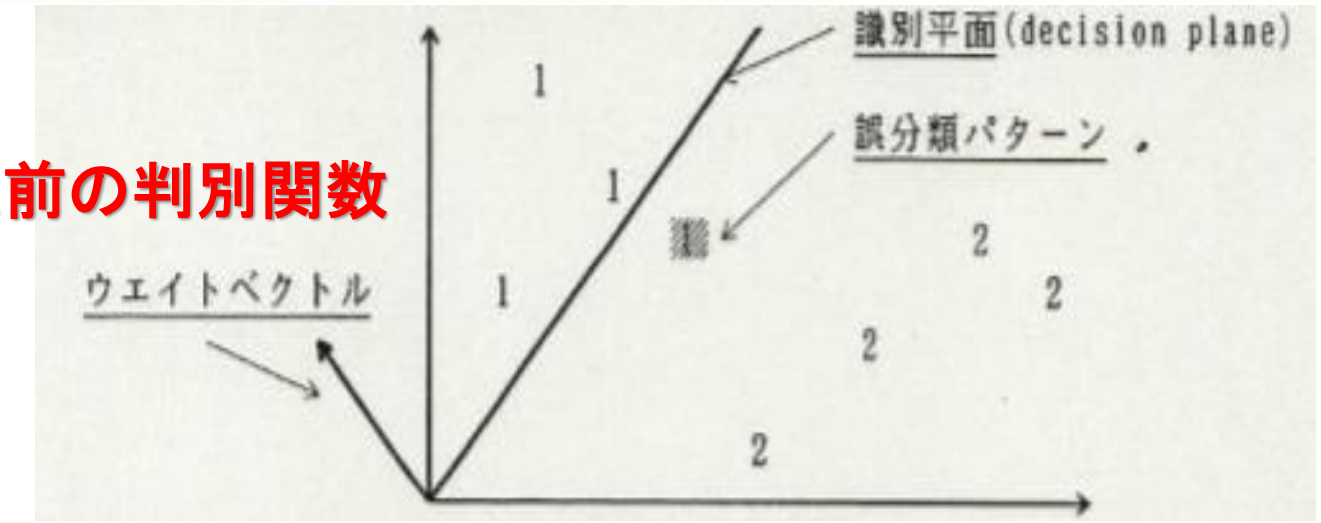
、（5）式のWnを（4）式に代入し、Cについて変換すると、

$$C = \frac{2(-Wo \cdot Xp)}{Xp \cdot Xp} \quad (6)$$

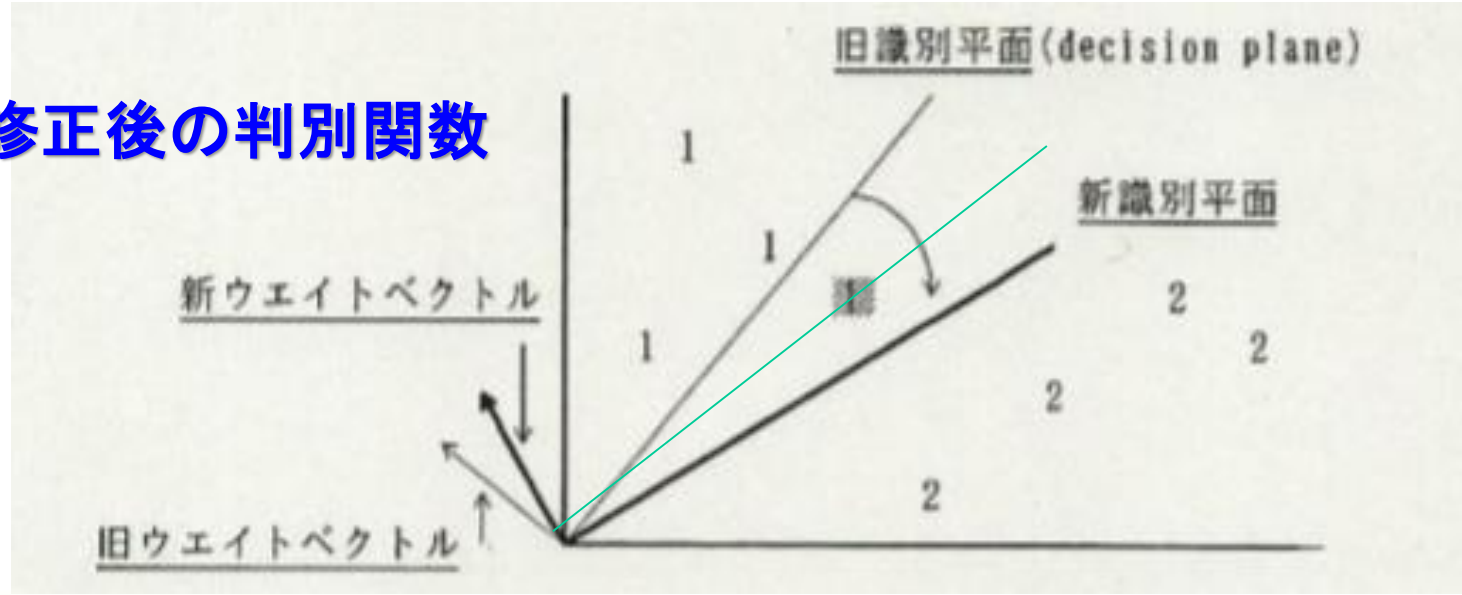
と（4）式から $D' = -Wo \cdot Xp$ 従って、

$$C = \frac{2D'}{Xp \cdot Xp} = -\frac{2D}{Xp \cdot Xp} \quad (7)$$

修正前の判別関数



修正後の判別関数



1. 人工知能概論と化学分野の人工知能

・ケモメトリックスと人工知能（ニューラルネットワーク型）の関係

各層の値の計算式ですが、(3) では各層の値を $\boldsymbol{l}^{(k)}$ で表しましたが、この例では 3 層ですから、もっと簡単に、
入力層を $\boldsymbol{x} = (x_1, x_2)$、隠れ層を $\boldsymbol{y} = (y_1, y_2, y_3)$、出力層を $\boldsymbol{z} = (z_1, z_2)$ と表すことにしましょう。

$$\boldsymbol{x} = (x_1, x_2) = \{(0,0), (0,1), (1,0), (1,1)\} \tag{7}$$

$$\begin{aligned} \boldsymbol{I_y} = (I_{y_1}, I_{y_2}, I_{y_3}) &= \boldsymbol{W_y x} + \boldsymbol{b_y} \\ &\begin{cases} I_{y_1} = W_{y_{11}} x_1 + W_{y_{12}} x_2 + b_{y_1} \\ I_{y_2} = W_{y_{21}} x_1 + W_{y_{22}} x_2 + b_{y_2} \\ I_{y_3} = W_{y_{31}} x_1 + W_{y_{32}} x_2 + b_{y_3} \end{cases} \end{aligned} \tag{8}$$

$$\boldsymbol{y} = (y_1, y_2, y_3) = \sigma_y(\boldsymbol{I_y}) \tag{9}$$

$$\begin{aligned} \boldsymbol{I_z} = (I_{z_1}, I_{z_2}) &= \boldsymbol{W_z y} + \boldsymbol{b_z} \\ &\begin{cases} I_{z_1} = W_{z_{11}} y_1 + W_{z_{12}} y_2 + W_{z_{13}} y_3 + b_{z_1} \\ I_{z_2} = W_{z_{21}} y_1 + W_{z_{22}} y_2 + W_{z_{23}} y_3 + b_{z_2} \end{cases} \end{aligned} \tag{10}$$

$$\boldsymbol{z} = (z_1, z_2) = \sigma_z(\boldsymbol{I_z}) \tag{11}$$

ここからは連鎖律も当たり前のように使っていきますので、(1)、(7) ～ (11) とにらめっこしながら、これからの式の導出をチェックしてみてください。

まずは z_1 のバイアスの偏微分から求めていきます。

$$\frac{\partial L}{\partial b_{z_1}} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial I_{z_1}} \cdot \frac{\partial I_{z_1}}{\partial b_{z_1}} = \frac{\partial L}{\partial z_1} \cdot \sigma_z'(I_{z_1}) \quad \left(\because \frac{\partial I_{z_1}}{\partial b_{z_1}} = 1 \right) \tag{12}$$

損失関数の偏微分 $(\partial L / \partial z_1)$ と活性化関数の偏微分 $(\partial z_1 / \partial I_{z_1} = \sigma_z'(I_{z_1}))$ が必要ですが、いまは両方とも明らかではありませんので、現時点でこれ以上は簡単にできません。
そしてここで、$\frac{\partial L}{\partial b_{z_1}} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial I_{z_1}}$ であることを利用して、

$$\frac{\partial L}{\partial W_{z_{11}}} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial I_{z_1}} \cdot \frac{\partial I_{z_1}}{\partial W_{z_{11}}} = \frac{\partial L}{\partial b_{z_1}} \cdot y_1 \tag{13}$$

$$\frac{\partial L}{\partial W_{z_{12}}} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial I_{z_1}} \cdot \frac{\partial I_{z_1}}{\partial W_{z_{12}}} = \frac{\partial L}{\partial b_{z_1}} \cdot y_2 \tag{14}$$

$$\frac{\partial L}{\partial W_{z_{13}}} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial I_{z_1}} \cdot \frac{\partial I_{z_1}}{\partial W_{z_{13}}} = \frac{\partial L}{\partial b_{z_1}} \cdot y_3 \tag{15}$$

と、バイアスの偏微分を使って重みの偏微分を表すことができます。z_2 のほうも同様で、

$$\frac{\partial L}{\partial b_{z_2}} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial I_{z_2}} \cdot \frac{\partial I_{z_2}}{\partial b_{z_2}} = \frac{\partial L}{\partial z_2} \cdot \sigma_z'(I_{z_2}) \quad \left(\because \frac{\partial I_{z_2}}{\partial b_{z_2}} = 1 \right) \tag{16}$$

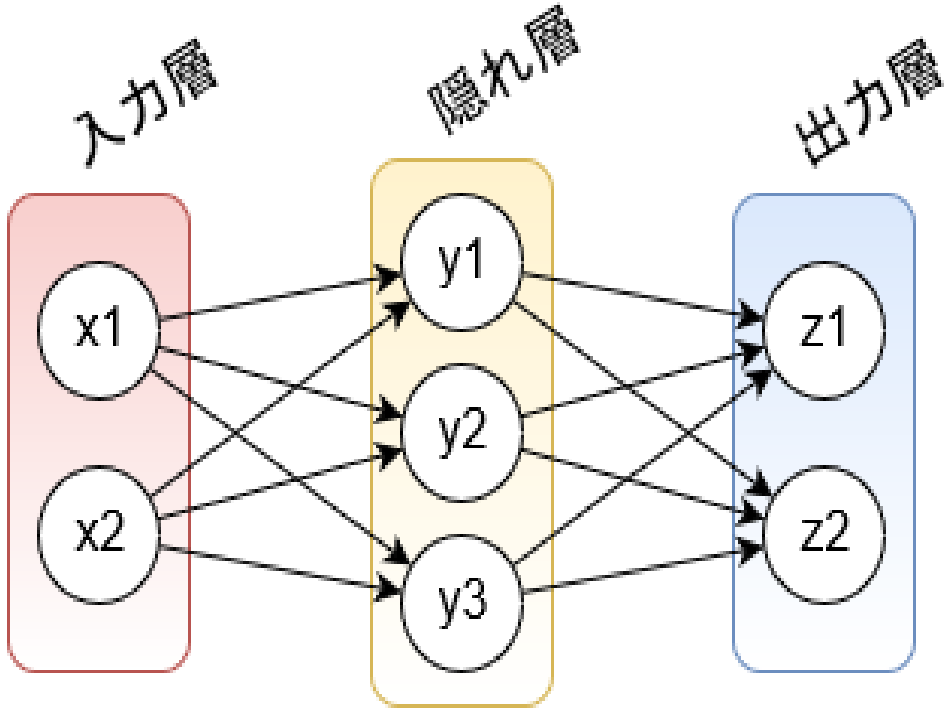
$$\frac{\partial L}{\partial W_{z_{21}}} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial I_{z_2}} \cdot \frac{\partial I_{z_2}}{\partial W_{z_{21}}} = \frac{\partial L}{\partial b_{z_2}} \cdot y_1 \tag{17}$$

$$\frac{\partial L}{\partial W_{z_{22}}} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial I_{z_2}} \cdot \frac{\partial I_{z_2}}{\partial W_{z_{22}}} = \frac{\partial L}{\partial b_{z_2}} \cdot y_2 \tag{18}$$

$$\frac{\partial L}{\partial W_{z_{23}}} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial I_{z_2}} \cdot \frac{\partial I_{z_2}}{\partial W_{z_{23}}} = \frac{\partial L}{\partial b_{z_2}} \cdot y_3 \tag{19}$$

となります。

バックプロパゲーション計算式の一部



https://qiita.com/43x2/items/50b55623c890564f1893

1. 人工知能概論と化学分野の人工知能
・ケモメトリックスと人工知能の関係

In Silico Data
Miracles by the KY-methods

多変量解析／パターン認識		人工知能
二クラス分類	⟷	犬と猫を識別する
判別関数の修正	⟷	識別のための学習
様々な判別手法	⟷	バックプロパゲーション

ニューラルネットワーク　　ニューラルネットワーク

Bayes、SVM、AdaBoost、他

1. 人工知能概論と化学分野の人工知能
・ケモメトリックスと人工知能の関係

多変量解析／パターン認識　　**人工知能**

機械学習

エラーフィードバックトレーニング　　学習

ニューラルネットワーク　⟷　ニューラルネットワーク

脳のシミュレーター　　脳のシミュレーター

分類手法

エラーフィードバックトレーニング、他

Bayes、SVM、AdaBoost、他

深層学習

1. 人工知能概論と化学分野の人工知能

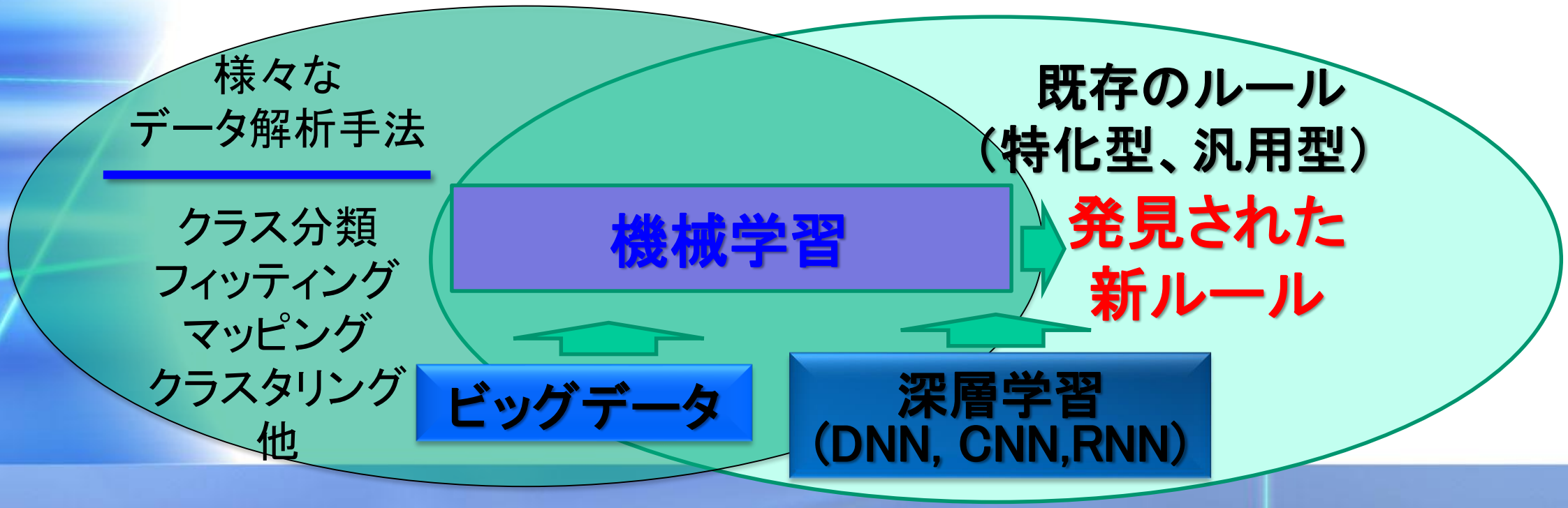
・ケモメトリックスと人工知能の関係

In Silico Data
Miracles by the KY-methods

現在における多変量解析/パターン認識と人工知能との関係

多変量解析/パターン認識と人工知能は
機械学習により繋がっている

多変量解析/パターン認識　　**人工知能**



2. ルールベース型人工知能

・特徴：人間のノウハウの活用

In Silico Data
Miracles by the KY-methods

ノウハウ取り出し＆活用型

人間の有するノウハウ

ルールとして取り出し

ノウハウデータベース

人工知能システム

人工知能の適用

ノウハウ発見＆活用型

開発関連情報
ビッグデータ

ニューラルネットワーク
深層学習

取り出されたノウハウや知識
ネットワーク構造として

人工知能の適用

2. ルールベース型人工知能

・特徴：人間のノウハウの活用

◇人工知能の基本的なアプローチ

初期の人工知能は人間の有するノウハウを取り出し、システム上で高度かつ高速に実施を目指す

適用限界：
- ルール取り出しの困難さ
- ルールのプログラミングでの限界
- ルールが多くいなると運用困難

ノウハウ**活用型**
人工知能

現在の人工知能は大量データから何らかのノウハウを取り出し、そのノウハウを用いて仕事をする

適用への期待：
- ルール取り出しの自動化

運用上の難点：
- ルールの取り出し困難
- 学習用サンプル数が大

ノウハウ**発見型**
人工知能

□化合物関連分野での様々な化学データサイエンス適用事例

In Silico Data
Miracles by the KY-methods

化合物関連データベース

パブリックデータベース
インハウスデータベース
研究用データベース

データベース融合、データ収集、
串刺し検索、
＊一元一項対応＊
＊プロトコル統一＊
＊化合物互変異性対応＊
＊サンプルポピュレーション＊
＊その他＊

データ解析と要因解析

ニクラス分類
重回帰（フィッテイング）
ニューラルネットワーク
マッピング
クラスタリング
PCA
PLS
グラフ表示
その他

＊チャンスコリレーション＊
＊オーバーフィッテイング＊
＊内挿性、外挿性＊
＊最小化合物数＊
＊線形、非線形＊

解析目的

構造－活性相関
ドラグデザイン
バーチャルスクリーニング
ドラグリポジショニング
リード化合物検索
リード化合物再構築
並列創薬

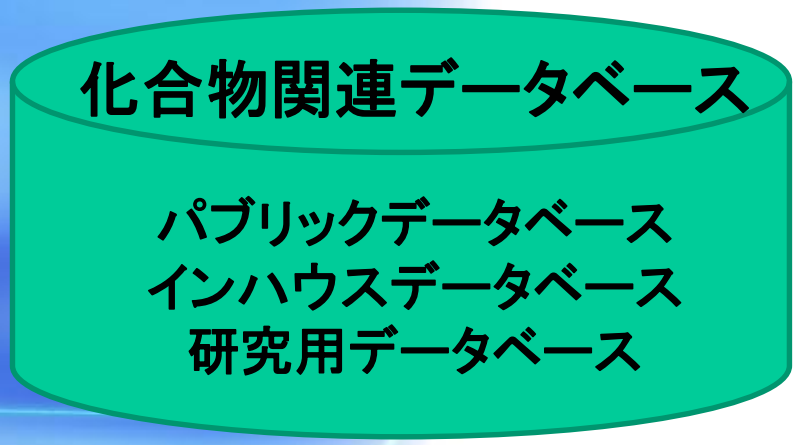
構造－毒性相関
化合物毒性評価/予測
脱毒性デザイン

構造－物性相関
機能性化合物デザイン

メタボロミクス

□化合物関連分野での様々な化学データサイエンス適用事例

・創薬関連（インシリコスクリーニング、AI創薬、ドラグリポジショニング）



薬理活性／毒性関連化合物データベース

薬理活性評価

判別関数

重回帰式

薬理活性／毒性がターゲット

解析目的

構造－活性相関
ドラグデザイン
バーチャルスクリーニング
ドラグリポジショニング
リード化合物検索
リード化合物再構築
並列創薬

構造－毒性相関
化合物毒性評価/予測
脱毒性デザイン

構造－物性相関
機能性化合物デザイン

□ルールベース型人工知能

・ルールベース型人工知能の特徴と限界
　　人間の考えるノウハウをシステム上でそのまま再現できるかの限界がある

例：簡単にルール化できる場合
①ケトンがあると活性が上昇
②窒素原子があると水溶性が上昇

例：様々な理由でルール化が困難な場合
①ケトンとアミン間でアロステリック効果があると活性が向上
　　⇒アロステリック効果がコード化困難な概念情報
②ラクトンがあると活性向上⇒環員数が定まらない不定情報

□ルールベース型人工知能

・ルールベース型人工知能の特徴と限界
　ノウハウルールの衝突が起こらないようにする

1．複数のルールが互いに入れ子関係にある時

①「ケトンがあると活性向上」
　⇒「アミドがあると活性低下」
　⇒「5および6員環ラクトンがあると活性低下」
②「12員環があると溶解性低下」
　⇒「縮合5員環があると溶解性変化なし」

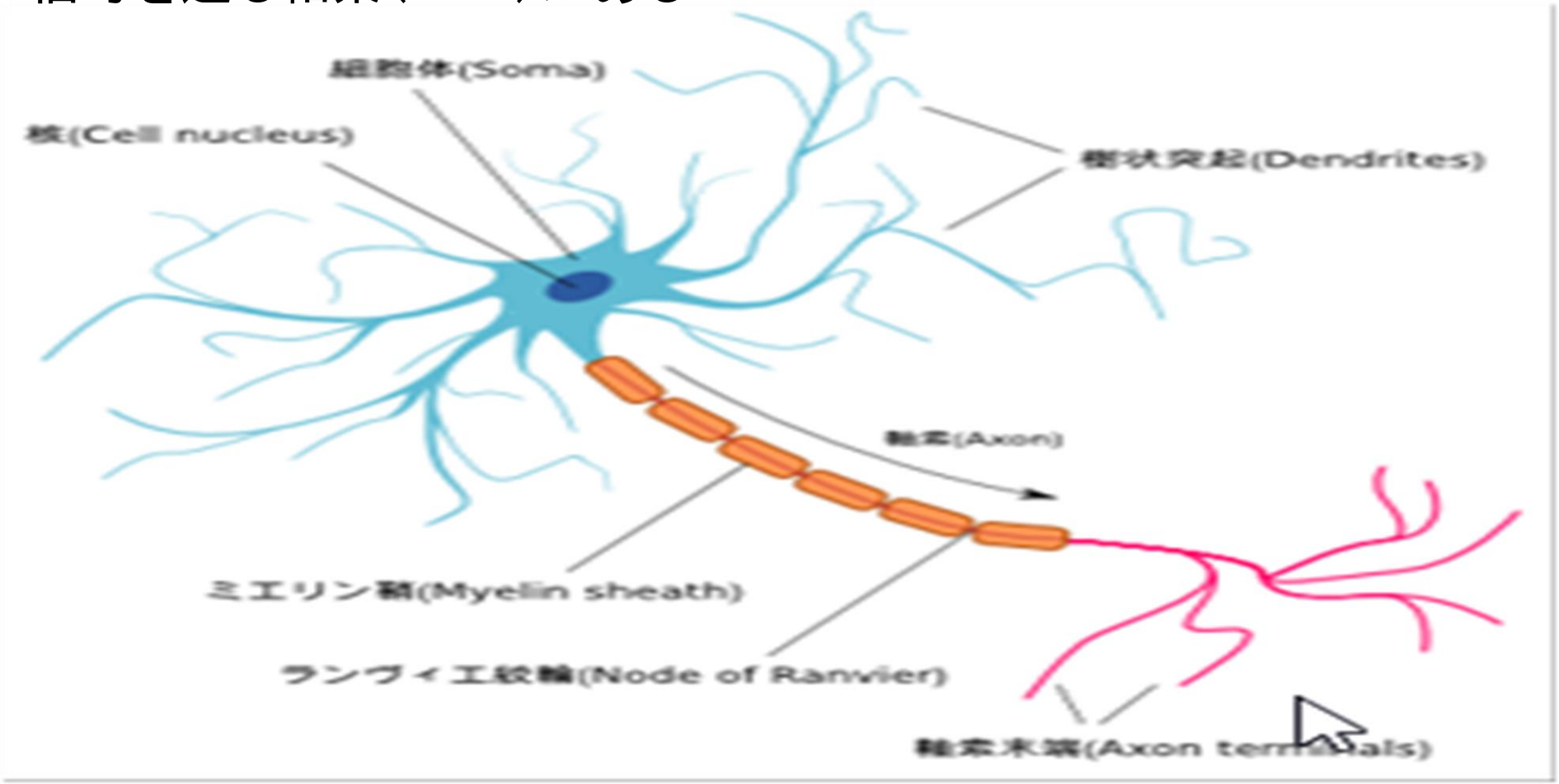
2．互いに矛盾するルールが入っている時

①「アルファハイドロキシケトンがあると活性向上」
　⇒「アルファベータ不飽和ケトンがあると活性低下」

□機械学習型人工知能

・特徴：ニューロンを用いた脳のシミュレーション

◇ニューロンは、他のニューロンからの信号を受ける樹状突起（dendrite）と、
　他のニューロンに信号を送る軸索（axon）がある

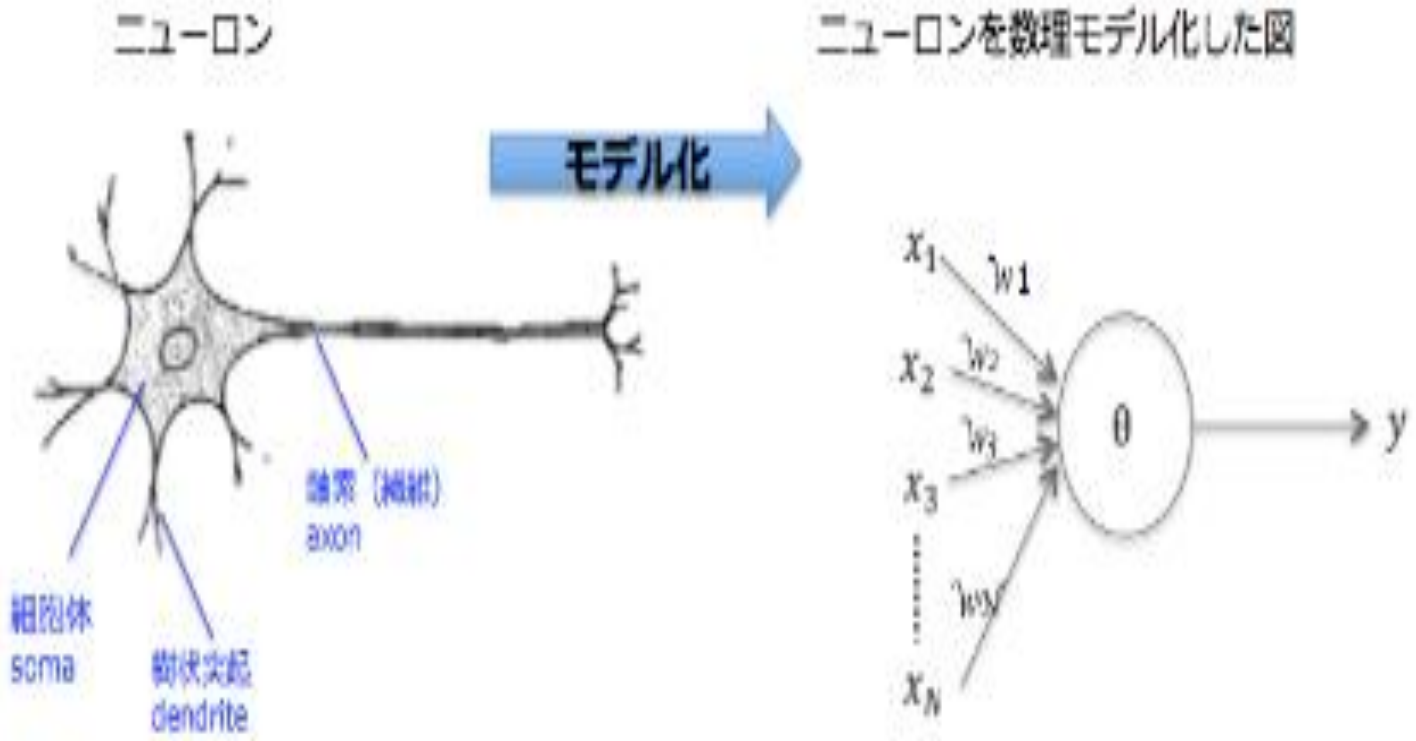


http://kazoo04.hatenablog.com/entry/agi-ac-3

□機械学習型人工知能

・特徴：ニューロンを用いた脳のシミュレーション

パーセプトロン

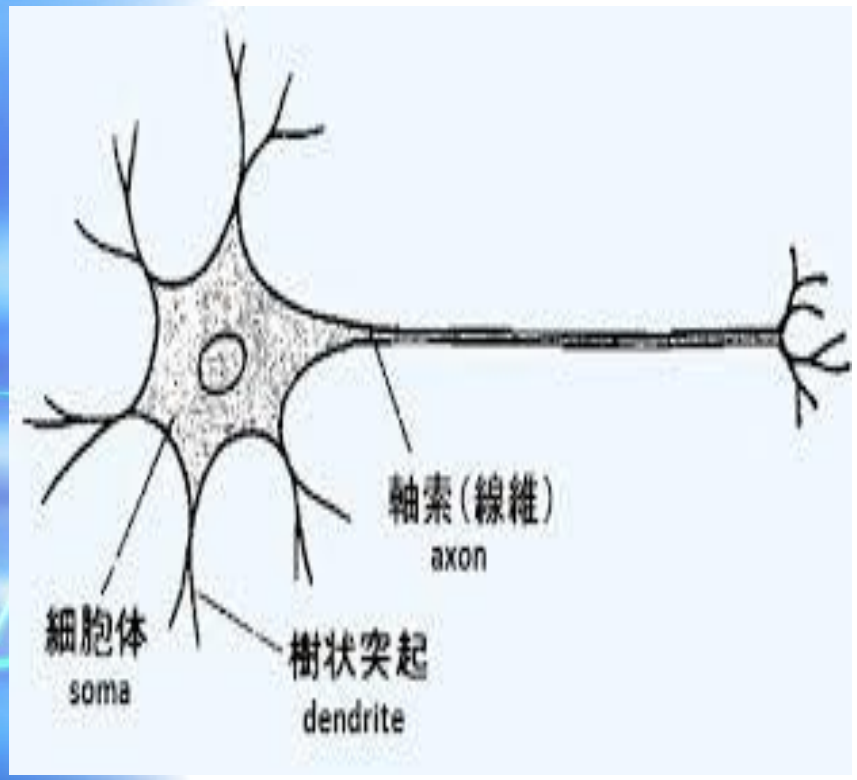


https://udemy.benesse.co.jp/ai/neural-network.html

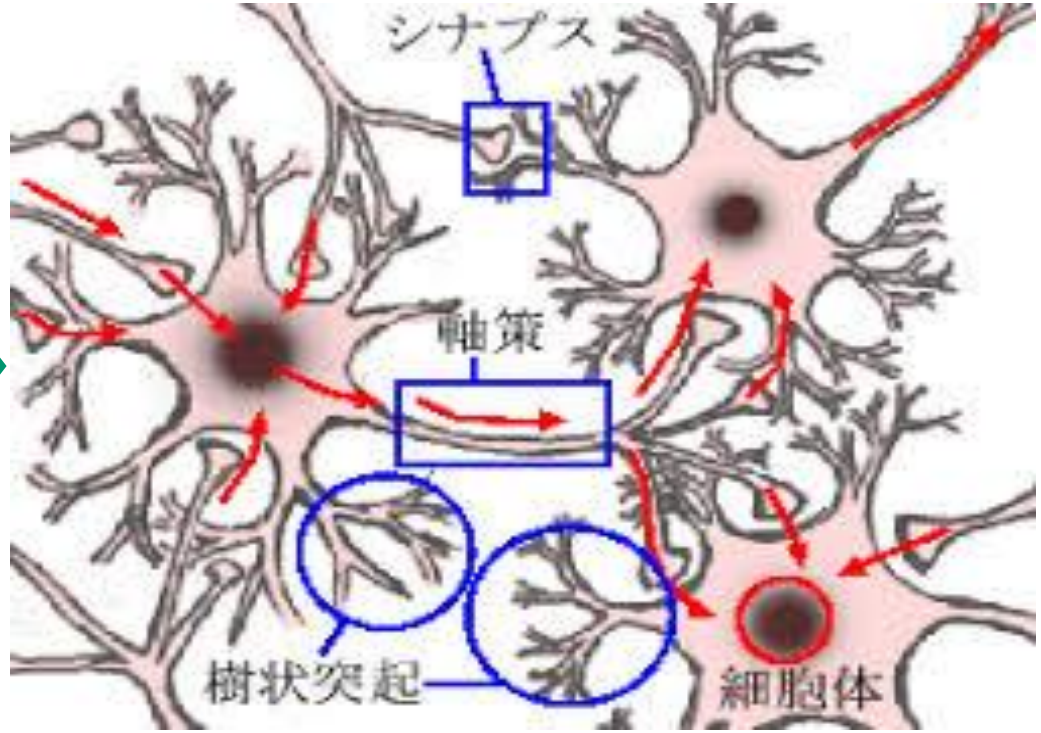
□機械学習型人工知能

・特徴：ニューロンを用いた脳のシミュレーション

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

□機械学習型人工知能

・特徴：ニューロンを用いた脳のシミュレーション

単ニューロンモデル

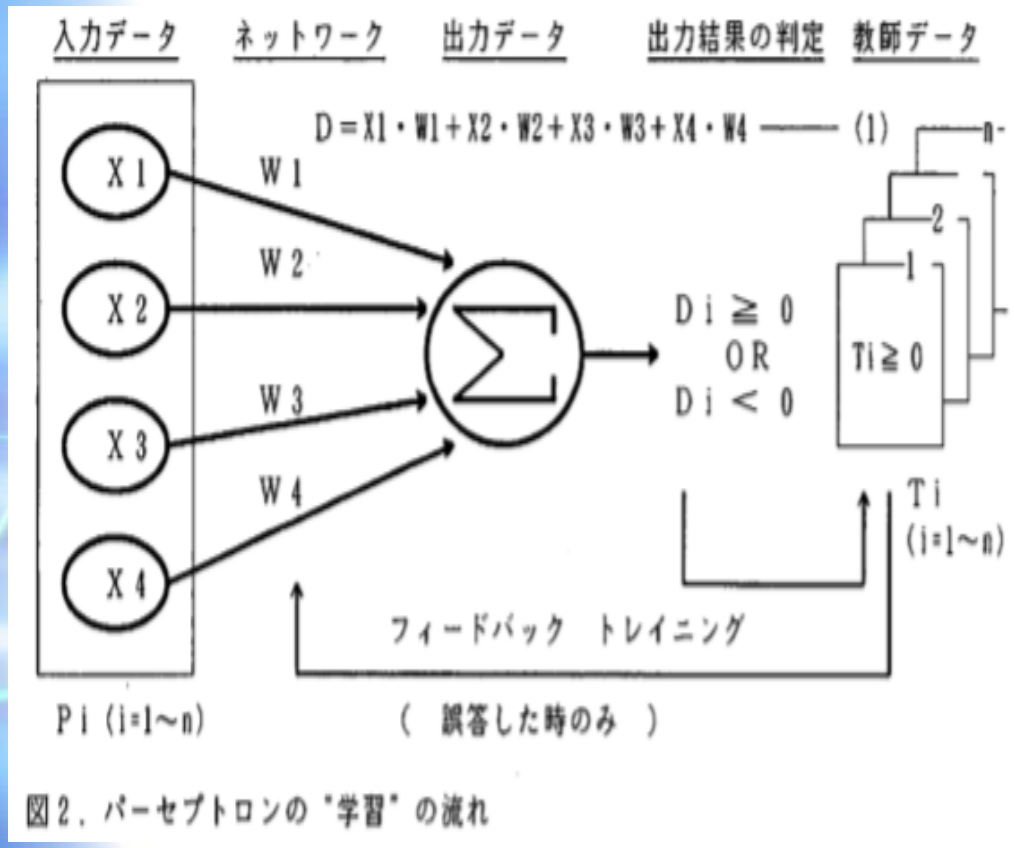
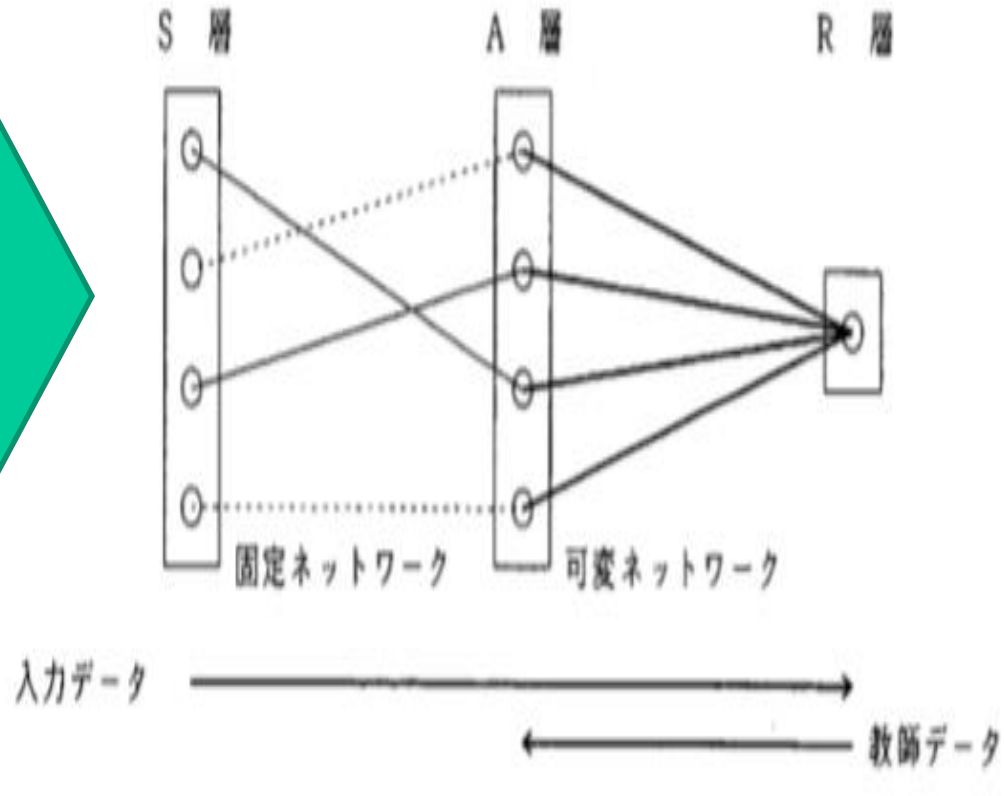


図2．パーセプトロンの"学習"の流れ

ネットワークニューロンモデル



パーセプトロン

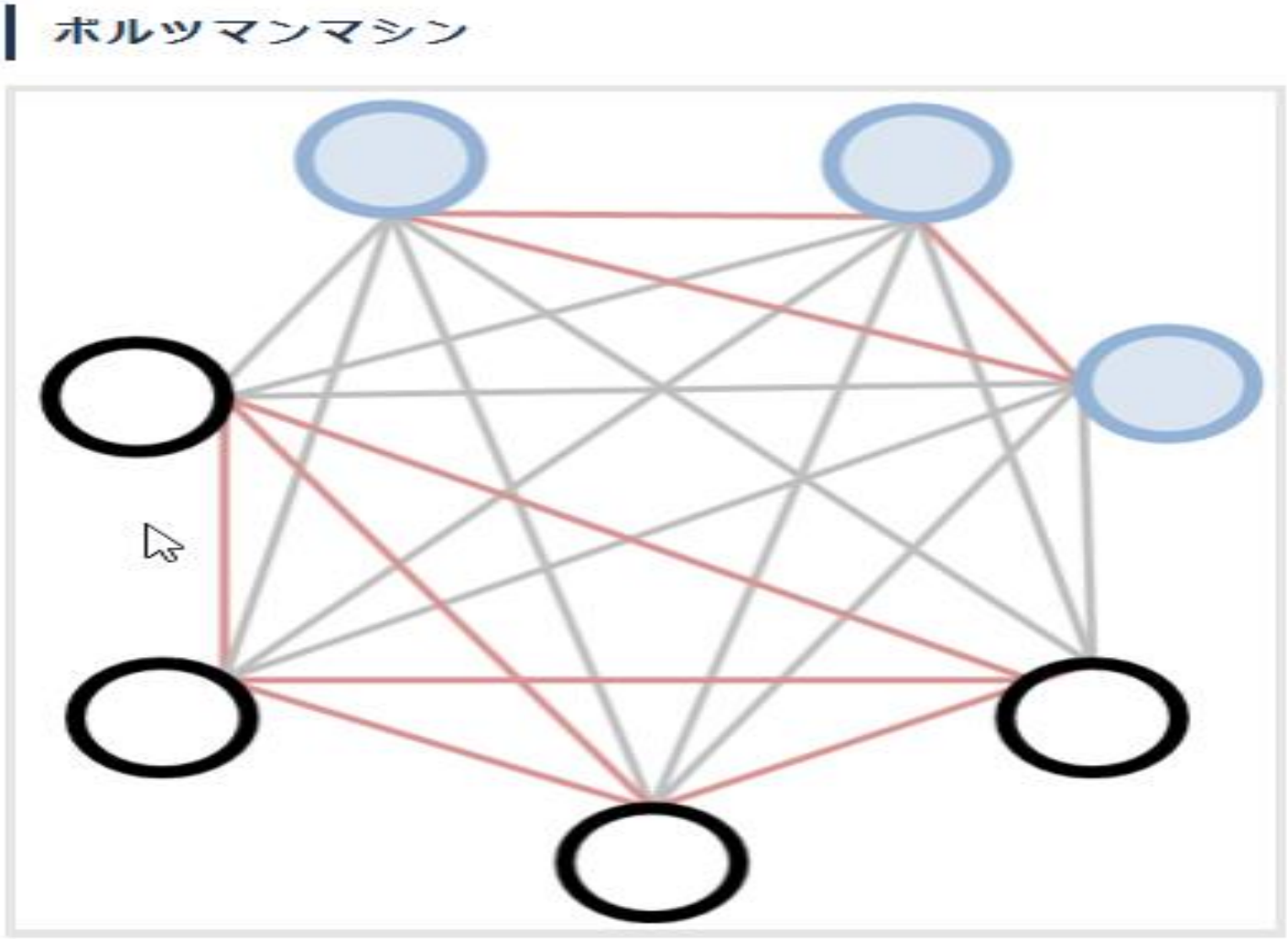
ニューラルネットワーク

□機械学習型人工知能

・特徴：ニューロンを用いた脳のシミュレーション

◇閉鎖型ニューラルネットワーク
ボルツマンマシン

最適化などに利用される



□機械学習型人工知能

・パーセプトロンとニューラルネットワーク

In Silico Data
Miracles by the KY-methods

＊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

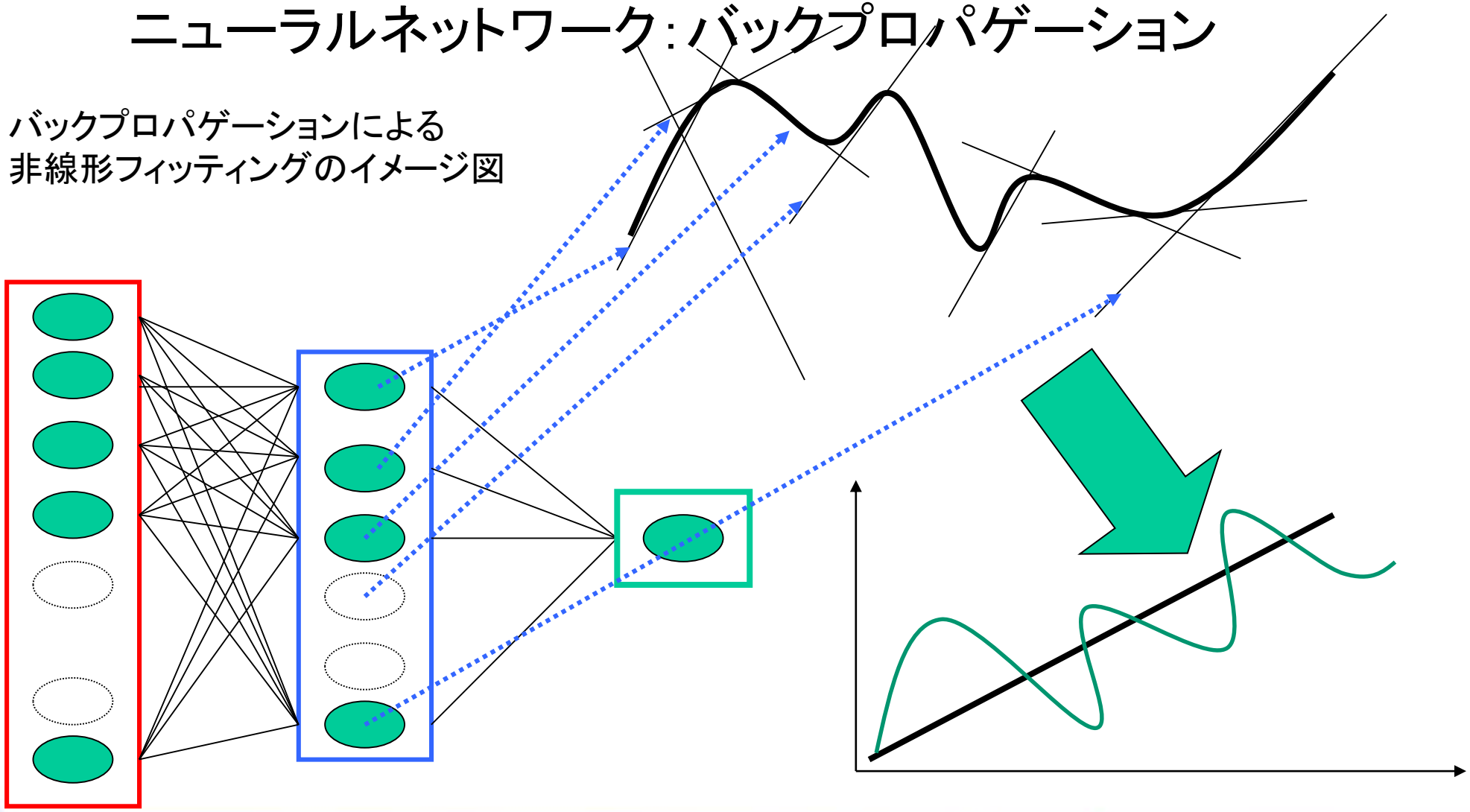
＊パーセプトロンは脳の機能を模したアプローチとして開発され、最終目標は
　AI(人工知能)への展開であった
＊ニューラルネットワークはその改良系で、ネットワーク構造が多層構造となった
＊最近の深層学習はニューラルネットワークのネットワーク構造をさらに複雑にした

□機械学習型人工知能
・パーセプトロンとニューラルネットワーク

In Silico Data
Miracles by the KY-methods

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
非線形フィッティングのイメージ図



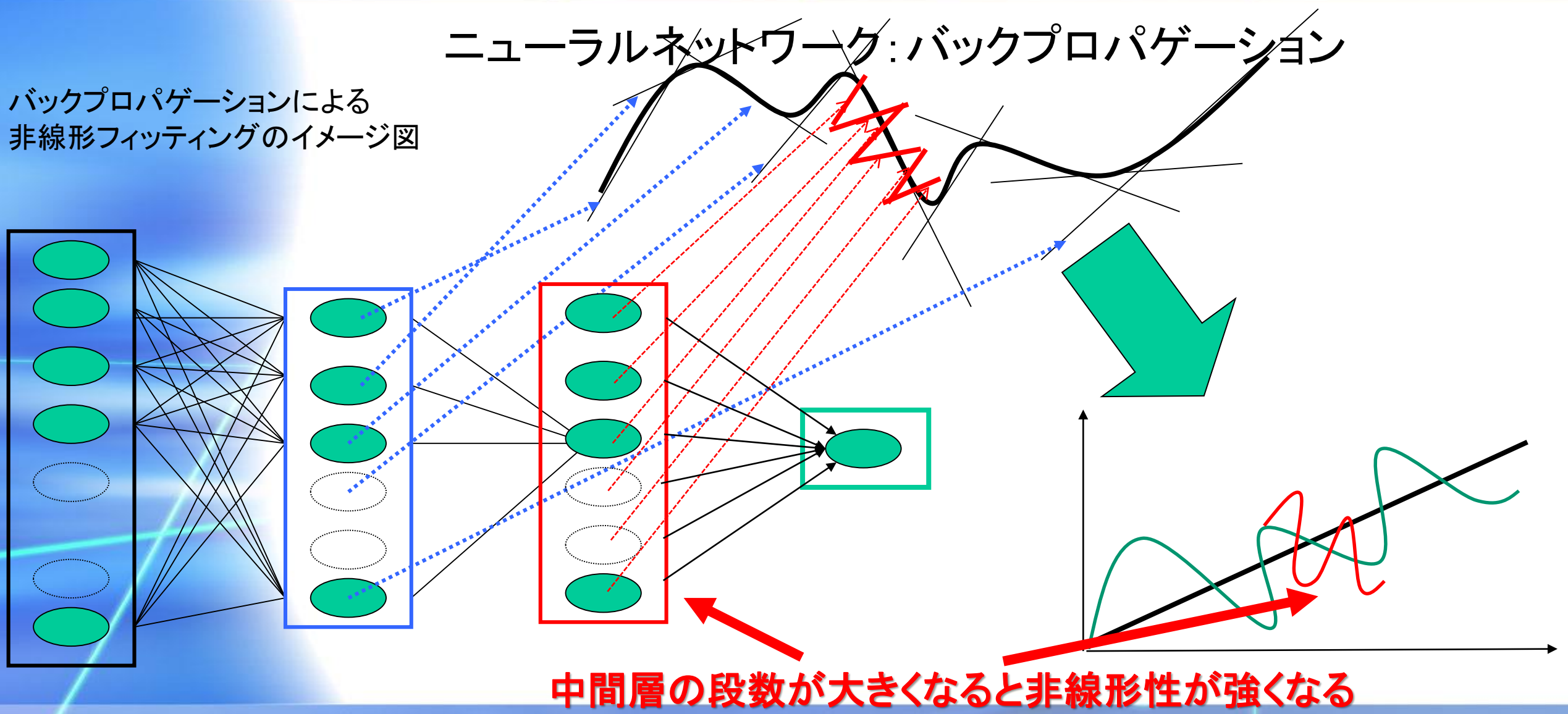
□機械学習型人工知能

・パーセプトロンとニューラルネットワーク

In Silico Data
Miracles by the KY-methods

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
非線形フィッティングのイメージ図



中間層の段数が大きくなると非線形性が強くなる

□機械学習型人工知能
・ニューラルネットワークのバックプロパゲーション

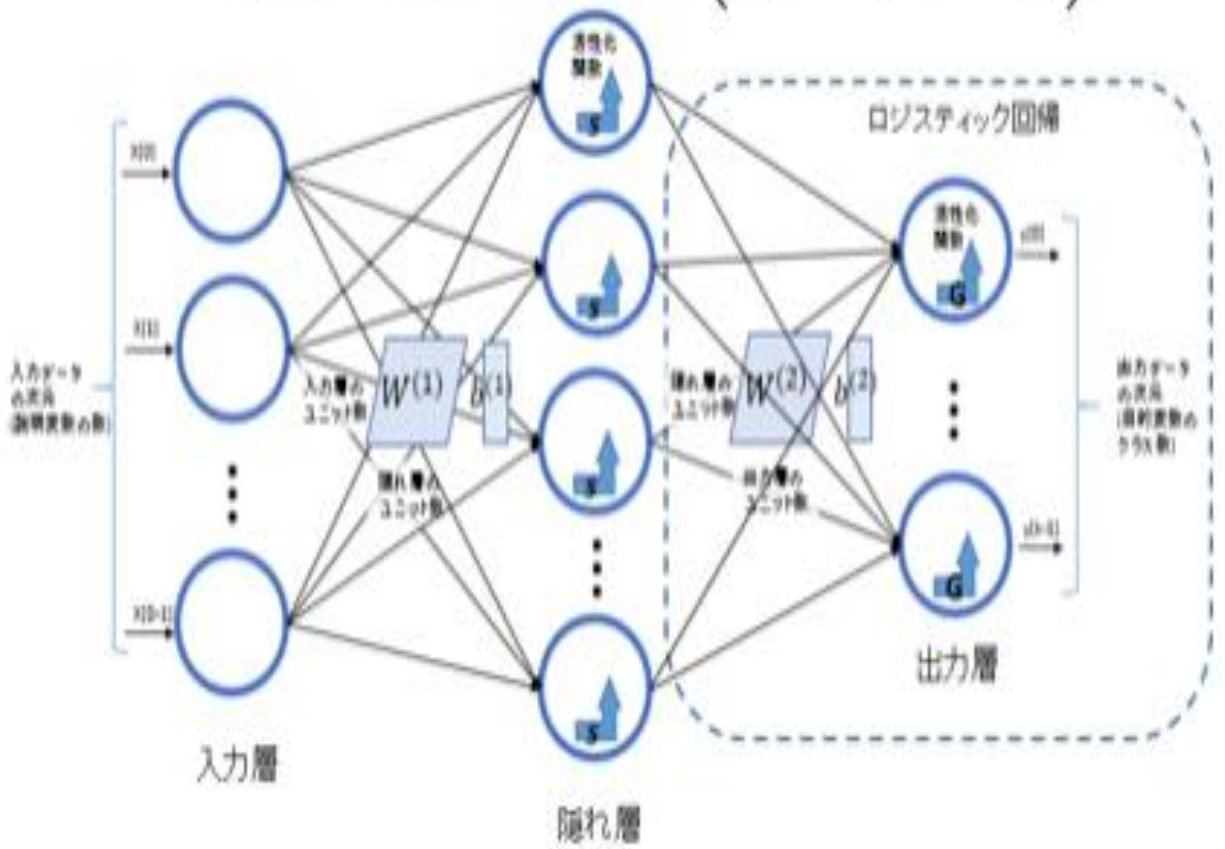
- $W^{(1)}$: 入力層 - 隠れ層の間で適用される係数行列。次元は 入力データの説明変数の数 D x 隠れ層のユニット数 D_h
- $b^{(1)}$: 入力層 - 隠れ層の間で適用される重みベクトル。次元は 隠れ層のユニット数 D_h
- s : 隠れ層の活性化関数。シグモイド関数もしくは $tanh$ (ハイパボリックタンジェント) をよく使う。
- $W^{(2)}$: 隠れ層 - 出力層の間で適用される係数行列 = ロジスティック回帰の係数行列。次元は 隠れ層のユニット数 x 出力データのクラス数
- $b^{(2)}$: 隠れ層 - 出力層の間で適用される重みベクトル = ロジスティック回帰の重みベクトル。次元は 出力データのクラス数
- G : 出力層の活性化関数。2クラスの場合は シグモイド、多クラスの場合はソフトマックス関数 (ロジスティック回帰と同じ)。

つまり 3層パーセプトロンでは、

1. 入力層 - 隠れ層: D 次元の入力を $W^{(1)}, b^{(1)}$ によって D_h 次元へと写像し、
2. 隠れ層 - 出力層: D_h 次元へと写像された入力を $W^{(2)}, b^{(2)}$ によってロジスティック回帰で学習 & クラス判別する

また、多層パーセプトロンの各層のユニット数 = その層に渡ってくるデータの次元と考えればよい。

$$f(x) = G(b^{(2)} + W^{(2)}\left(s(b^{(1)} + W^{(1)}x)\right))$$



http://sinhrks.hatenablog.com/entry/2014/11/30/085119

□機械学習型人工知能

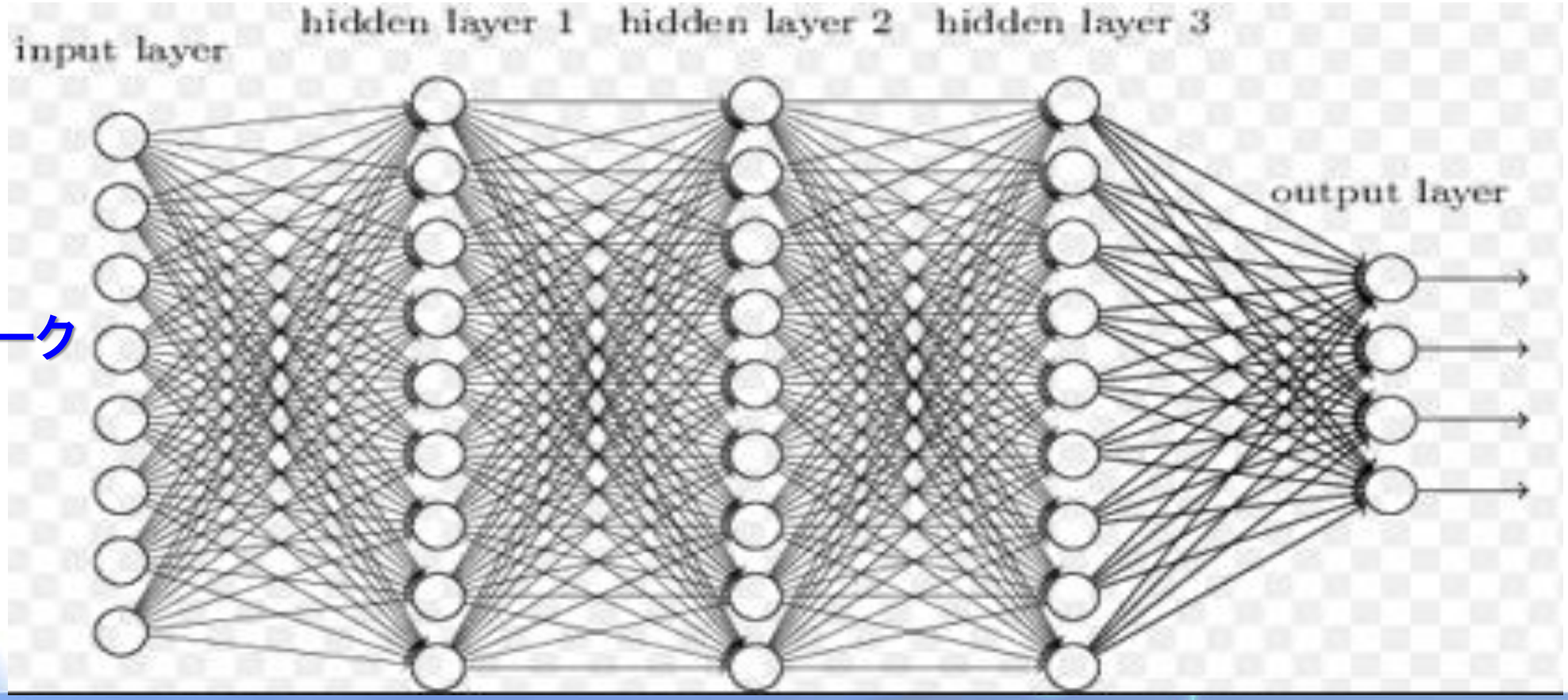
・実行環境：機械学習手法の発展（深層学習：ディープラーニング）

◇ニューラルネットワークから深層学習へ

中間層を増やすことで分類性が向上することは明白であったが、効率的なバックプロパゲーションアルゴリズムが無かったので展開が遅れていた。しかし、効率的に学習結果を前に戻すアルゴリズムが開発されたことで深層学習が開発された。

現在では、中間層が多層になった深層学習が次世代ニューラルネットワーク（人工知能）として脚光を浴びている。

順伝搬型
ニューラルネットワーク

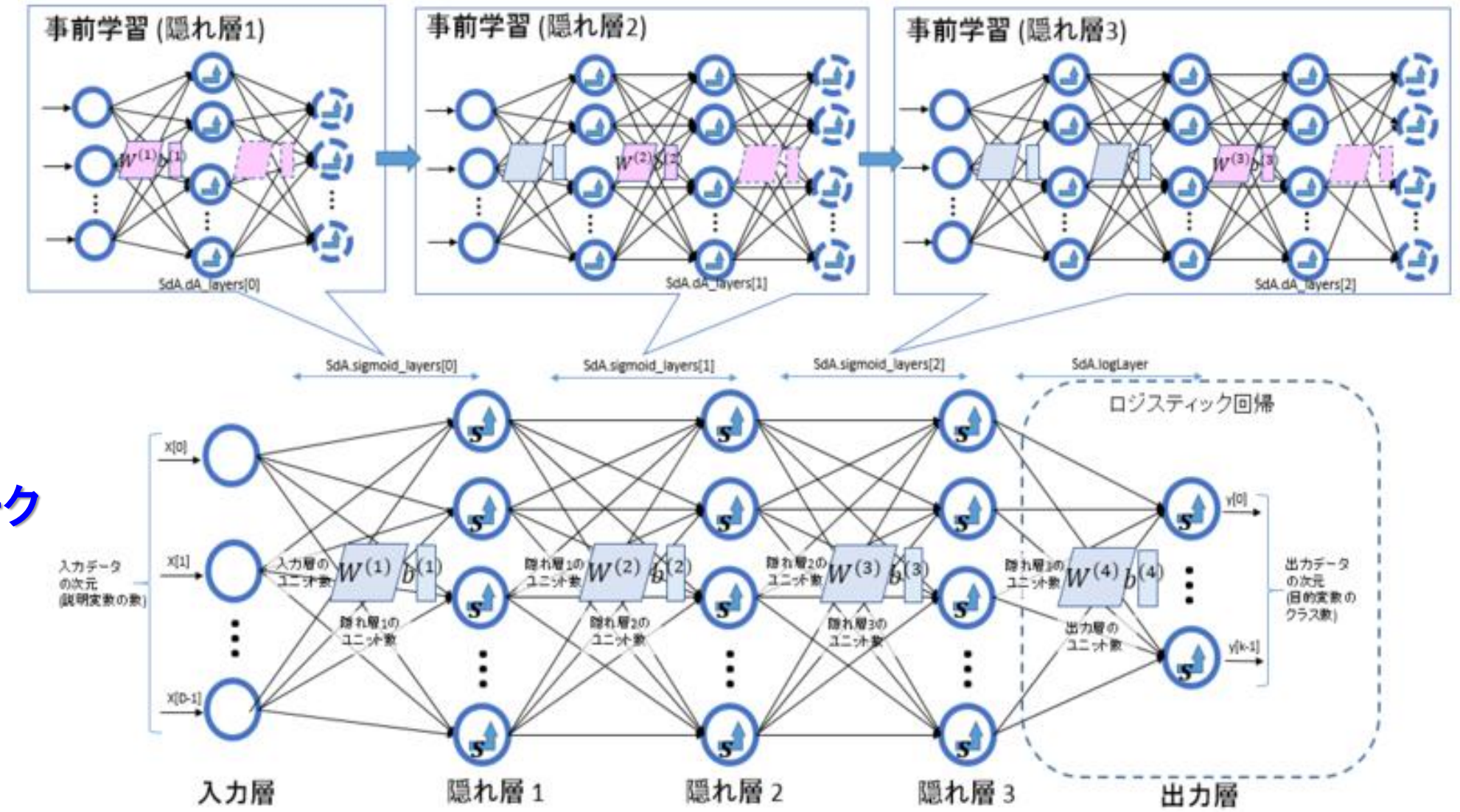


https://nnadl-ja.github.io/nnadl_site_ja/chap6.html

□機械学習型人工知能

・実行環境：機械学習手法の発展（深層学習：ディープラーニング）

順伝搬型
ニューラルネットワーク



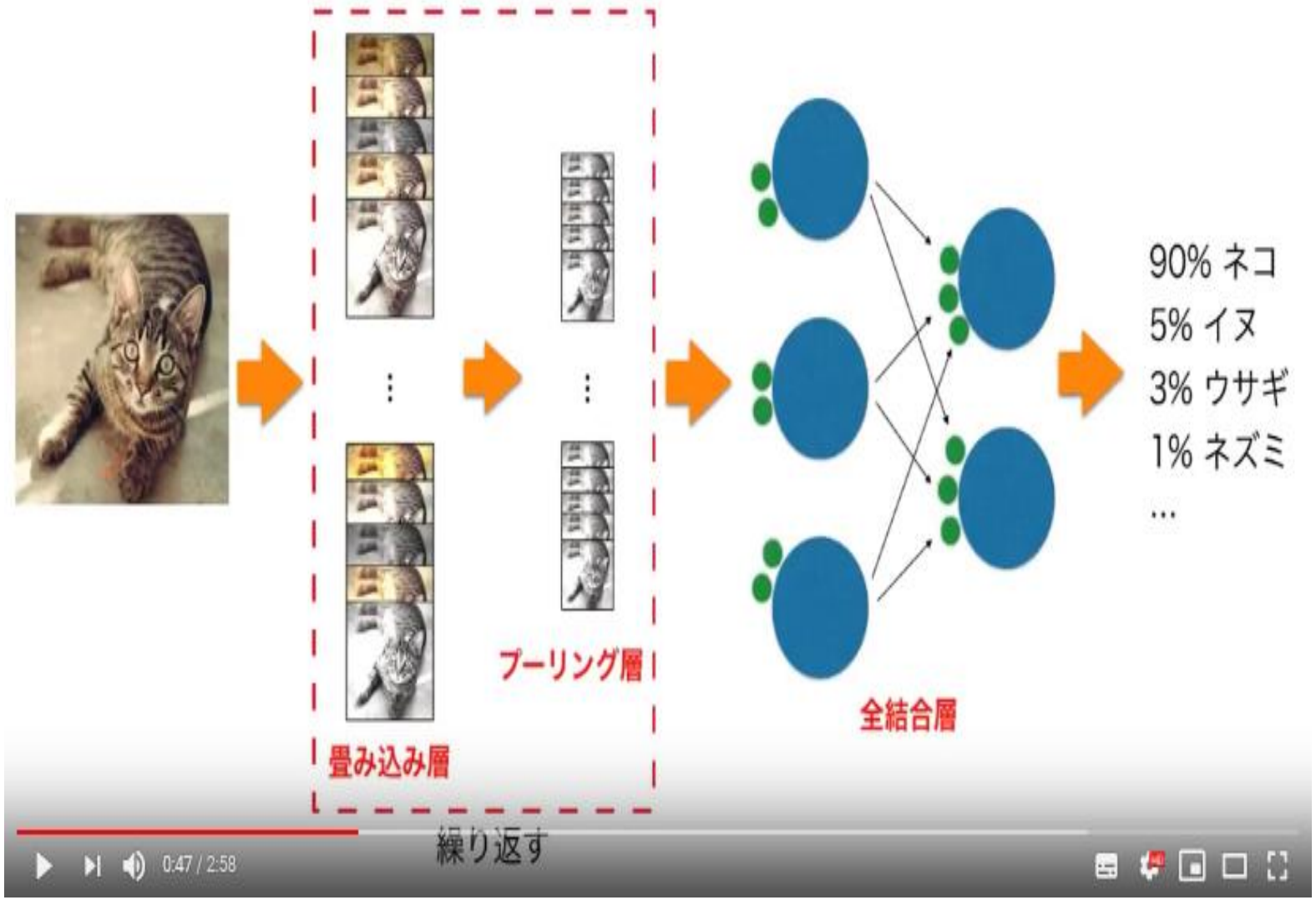
http://sinhrks.hatenablog.com/entry/2014/12/21/232145

□機械学習型人工知能

・開発歴：畳み込みニューラルネットワーク

畳み込みニューラルネットワーク
（Convolutional Neural Network）

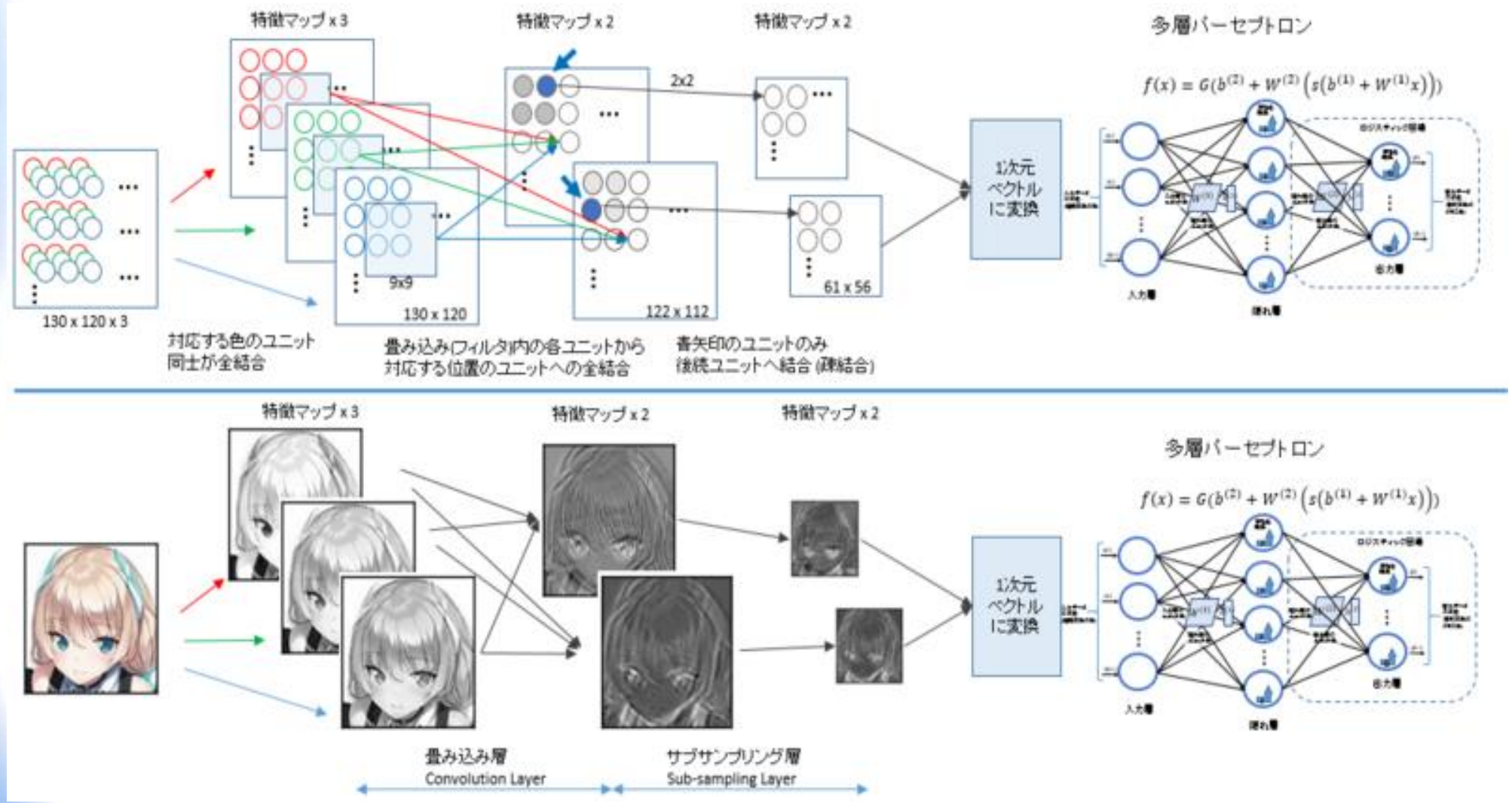
CNNは、全結合層だけでなく
畳み込み層(Convolution Layer)と
プーリング層(Pooling Layer)から
構成されるニューラルネットワークのことだ。



https://www.youtube.com/watch?v=Kr84vlulqTA

□機械学習型人工知能

・開発歴：畳み込みニューラルネットワーク



http://sinhrks.hatenablog.com/entry/2014/12/07/203048

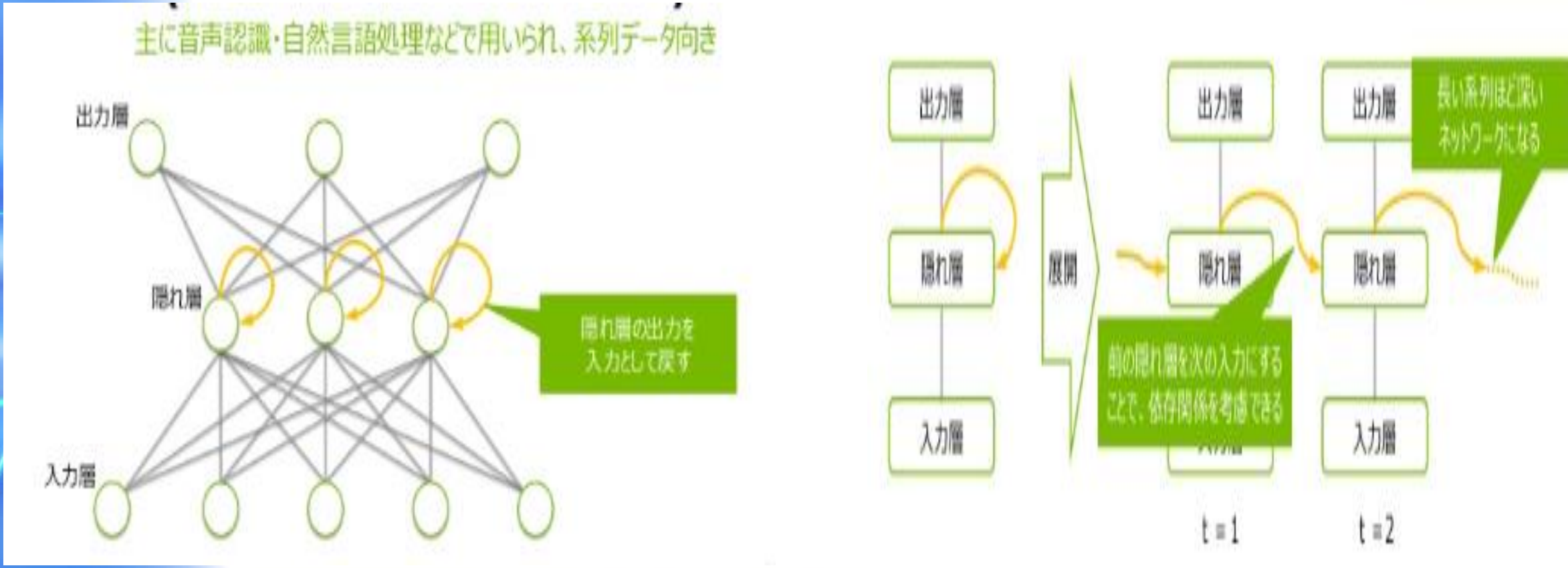
□機械学習型人工知能

・開発歴：リカレント（再帰型）ニューラルネットワーク（RNN）

◇リカレント（再帰方）ニューラルネットワーク（Recurrent Neural Networks）

・中間層の出力データを再び入力データとして利用することを特徴とする

・時系列データの解析や言語処理等に利用される



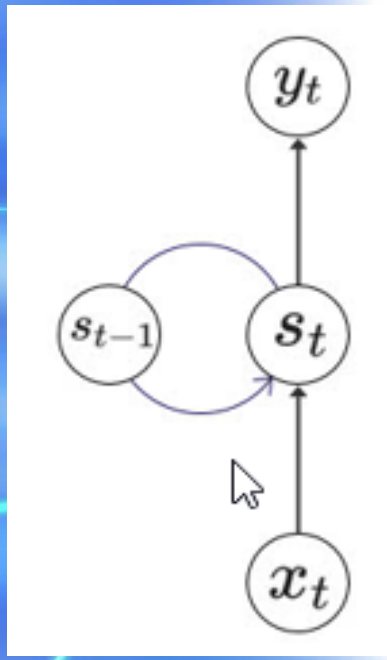
https://www.slideshare.net/NVIDIAJapan/ss-75109746

□機械学習型人工知能

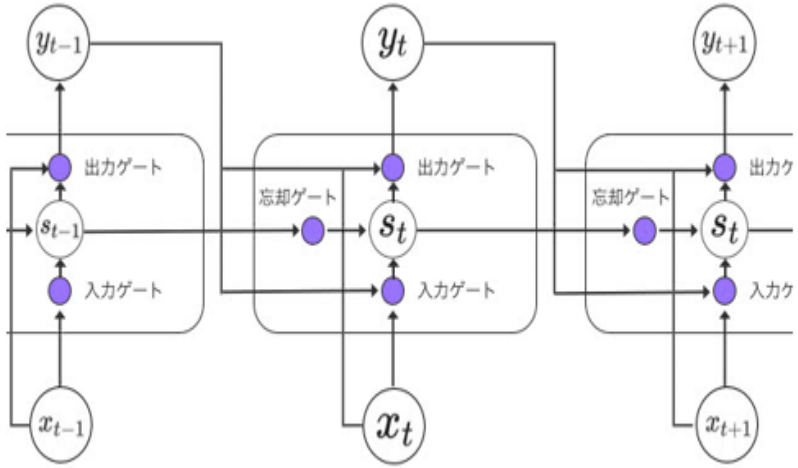
・開発歴：リカレントニューラルネットワーク（RNN）

◇リカレントニューラルネットワークには様々なタイプがある
　基本の考え方は、自身の出力を改めて入力に使うことである

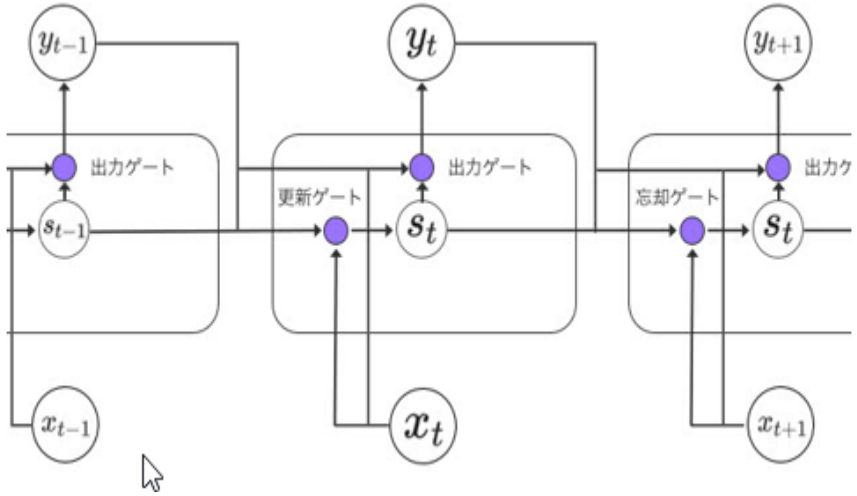
Simple RNN



LSTM（Long Short-Term Memory）



GRU（Gated Recurrent Unit）



https://deepage.net/deep_learning/2017/05/23/recurrent-neural-networks.html

□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

In Silico Data
Miracles by the KY-methods

◇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

In Silico Data
Miracles by the KY-methods

◇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

□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

In Silico Data
Miracles by the KY-methods

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

◆化学分野で人工知能を適用する時の注意とまとめ

In Silico Data
Miracles by the KY-methods

・機械学習型人工知能適用上での解決すべき点や対応

1. 化合物情報の扱い

・化合物の画像情報、異性体の扱い、一元一項対応

2. サンプル数の問題

・異性体の扱いによるサンプル数の水増し
・少ないサンプル時の対応；転移学習、スパースモデリング、他

3. 要因説明の問題

・ネットワークからの情報取り出し
・ネットワーク構造の単純化

化合物情報の扱い：　Graph convolution

Graph convolutionは最近のニューラルネットワーク型人工知能の展開過程で、化合物情報を数値データ化する手法としてトポロジー理論を用いて展開された。

トポロジーによる化合物情報の数値化であるが、同様のアプローチは構造－活性相関が展開された当初より既に展開されていた。

トポロジーによる化合物の数値化には、Hosoyaインデックス、MCI(Molecular Connectivity Index)、その他、多種多様のインデックスが提唱されて来た。

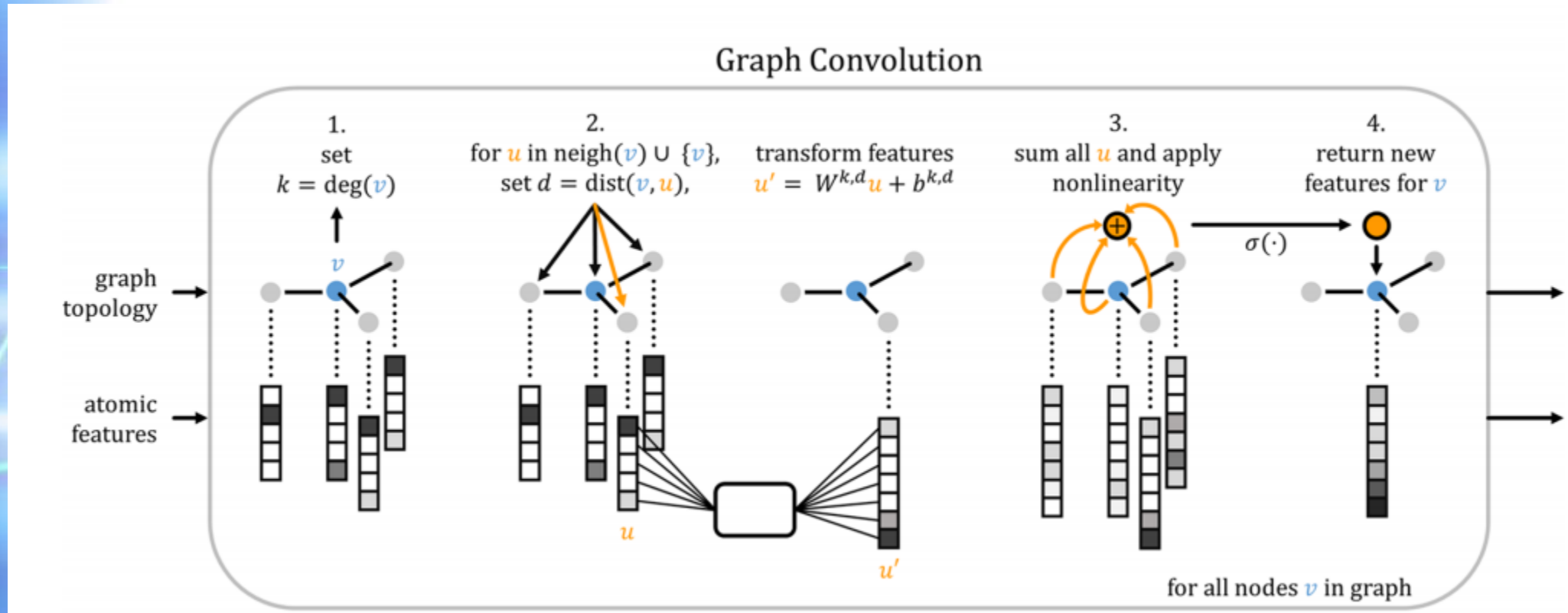
それぞれのインデックスは、創薬や物性等の研究分野で精力的に展開され、様々な薬理活性や、化合物物性との相関研究が発表されてきた。

◆化学分野で人工知能を適用する時の注意とまとめ

・機械学習型人工知能適用上での解決すべき点

In Silico Data
Miracles by the KY-methods

化合物情報の扱い：　Graph convolution



https://kivantium.net/deep-for-chem

□化学分野で人工知能を適用する時の注意とまとめ

・機械学習型人工知能適用上での解決すべき点

In Silico Data
Miracles by the KY-methods

化合物情報の扱い：　Graph convolution

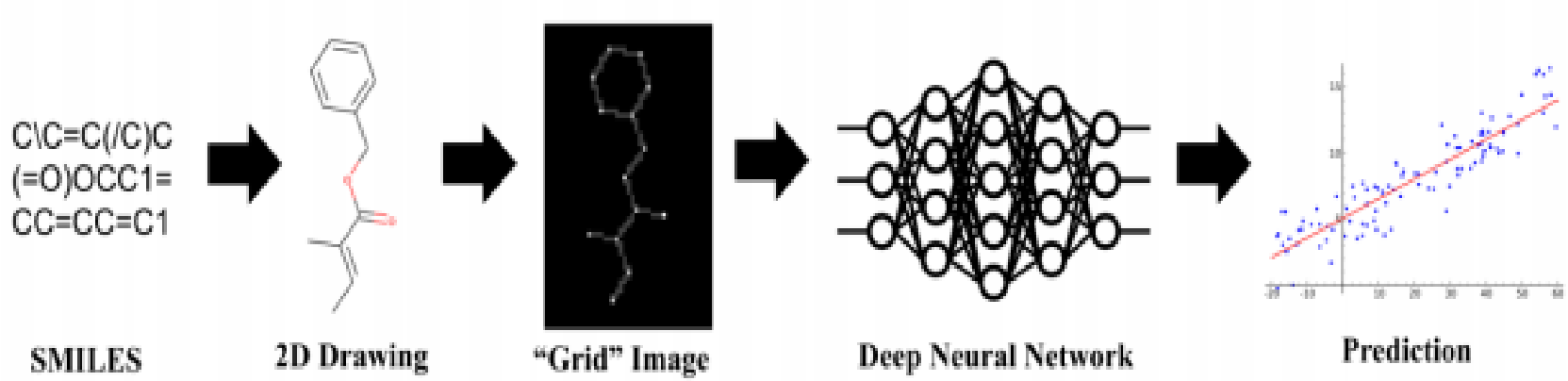


Figure 2: Illustration of the Chemception framework. After a SMILES to structure conversion, the 2D images are mapped onto an 80 x 80 image that serves as the input image data for training a deep neural network to predict toxicity, activity, and solvation properties.

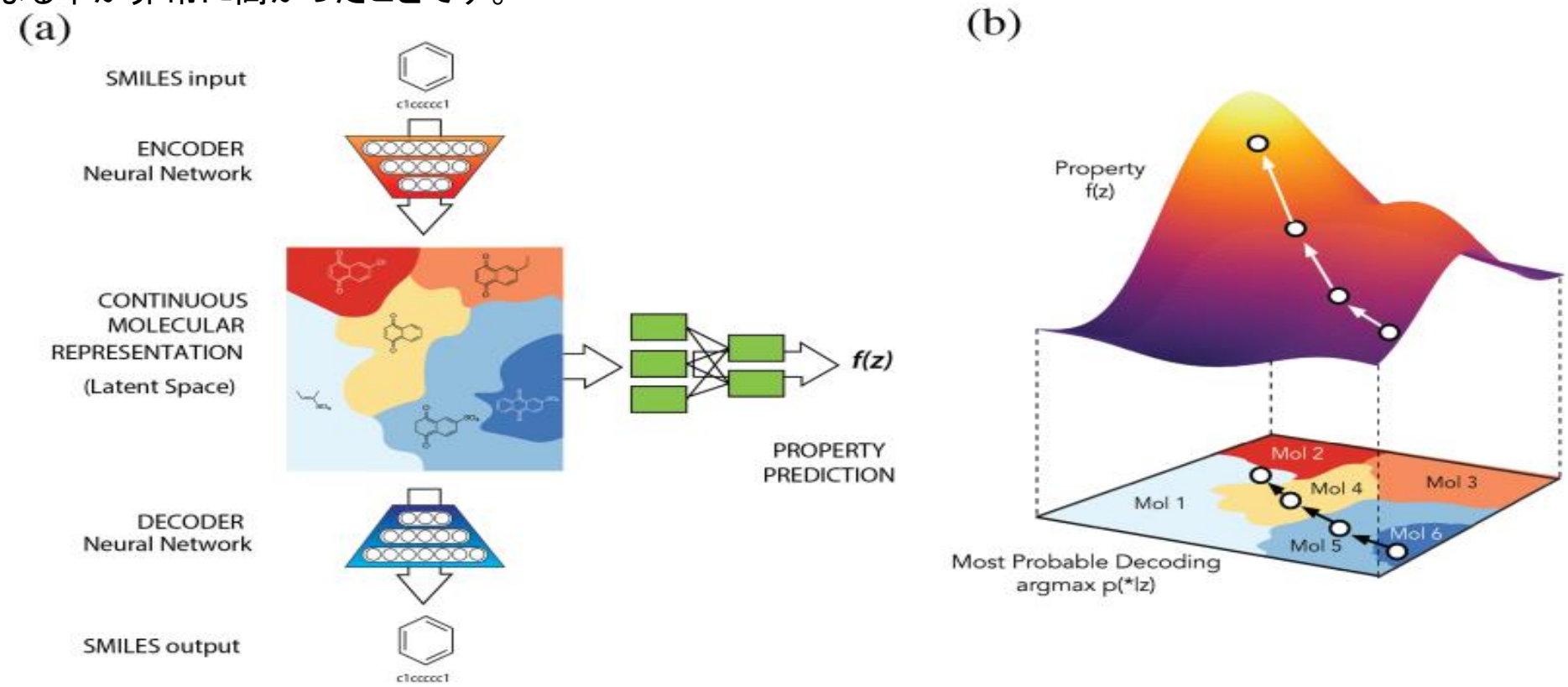
https://kivantium.net/deep-for-chem

□化学分野で人工知能を適用する時の注意とまとめ

・機械学習型人工知能適用上での解決すべき点

化合物情報の扱い： Graph convolution

分子設計にDeep Leaningを持ち込んだ研究が[Gómez-Bombarelli+, 2016]です。この研究では分子の文字列表現であるSMILES記法をvariational autoencoder (VAE) を用いて実数ベクトルに変換し、ベイズ最適化で最適化したベクトルをSMILESに戻すことで分子を設計しています。この手法の問題点はVAE空間上で最適化ベクトルをSMILESに戻したときに生成される文字列が文法的に正しくないなどの理由で分子と対応しなくなる率が非常に高かったことです。



https://kivantium.net/deep-for-chem

7. 化学分野で人工知能を適用する時の注意とまとめ
・機械学習型人工知能適用上での解決すべき点

In Silico Data
Miracles by the KY-methods

化合物情報の扱い： Graph convolution

文法的に正しくないSMILESの文字列が生成される問題を解決するために、VAEの入出力にSMILESの文字列をそのまま使うのではなくSMILESを生成する文脈自由文法の生成規則列を使うことにしたのが[Kusner+, 2017]のGrammar Variational Autoencoderです。この研究で技術的に面白いところはVAE表現から文字列を生成する際にプッシュダウンオートマトンを考えて、現在スタックの一番上にある文字から選択できない生成規則の確率を0にする工夫を導入しているところです。この工夫により生成される文字列はSMILESの文法的に正しいものに限定することができるためデコードの効率が上がるほか、潜在空間自体もよりよいものになったと主張されています。

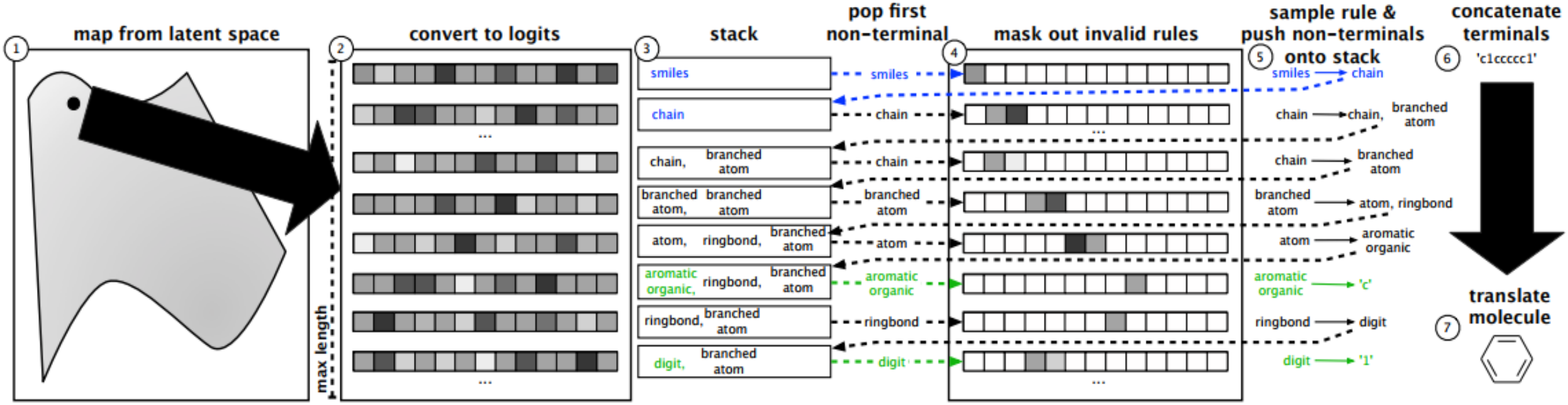


Figure 2. The decoder of the GVAE. See text for details.

https://kivantium.net/deep-for-chem

□化学分野で人工知能を適用する時の注意とまとめ

□現在の人工知能による医療や創薬へのアプローチ

- ●WATSONとの連携による医療関連アプローチ
- ●深層学習採用による創薬関連アプローチ

現在時点では話題提供的なものが多く、深層学習もトライアル的で、本格的な成功事例を伴うケースはないといえる。
多数のJournalを入力し、結果として新薬候補が得られたということも報告されているが、実際に薬理活性が出たという報告はない。

口化学分野で人工知能を適用する時の注意とまとめ

・機械学習型人工知能の有する基本的弱点

日本経済新聞　2018年12月9日（朝刊）
タイトル「**AI、弱点克服へ厚い壁**」

本文中：感情表現や判断力まだまだ

1.「僕の“耳”は遠くの音声を認識できない。マイクの近くで話してね」
2.「僕は状況に合わせて抑揚をつける話し方が下手なんだ。・・・人間のように喜怒哀楽をうまく表現できない」
3.「僕は賢いと思われているけど、物覚えがいいだけで、判断力は高くないよ」
4.「人間には考えられない見間違いを起こすこともあるよ」
5.「人間は手を使って様々な作業が器用にできるね。だけど僕は、本をつかむのさえ苦労しているよ」

図中：今のAIではできないことも多い

1. 人間には考えられない見間違いも
2. 離れた場所の音声の内容を認識できない
3. 膨大なデータがないと賢くならない
4. 状況に応じた抑揚をつけられない
5. 人間の何気ない動きで難しいものも

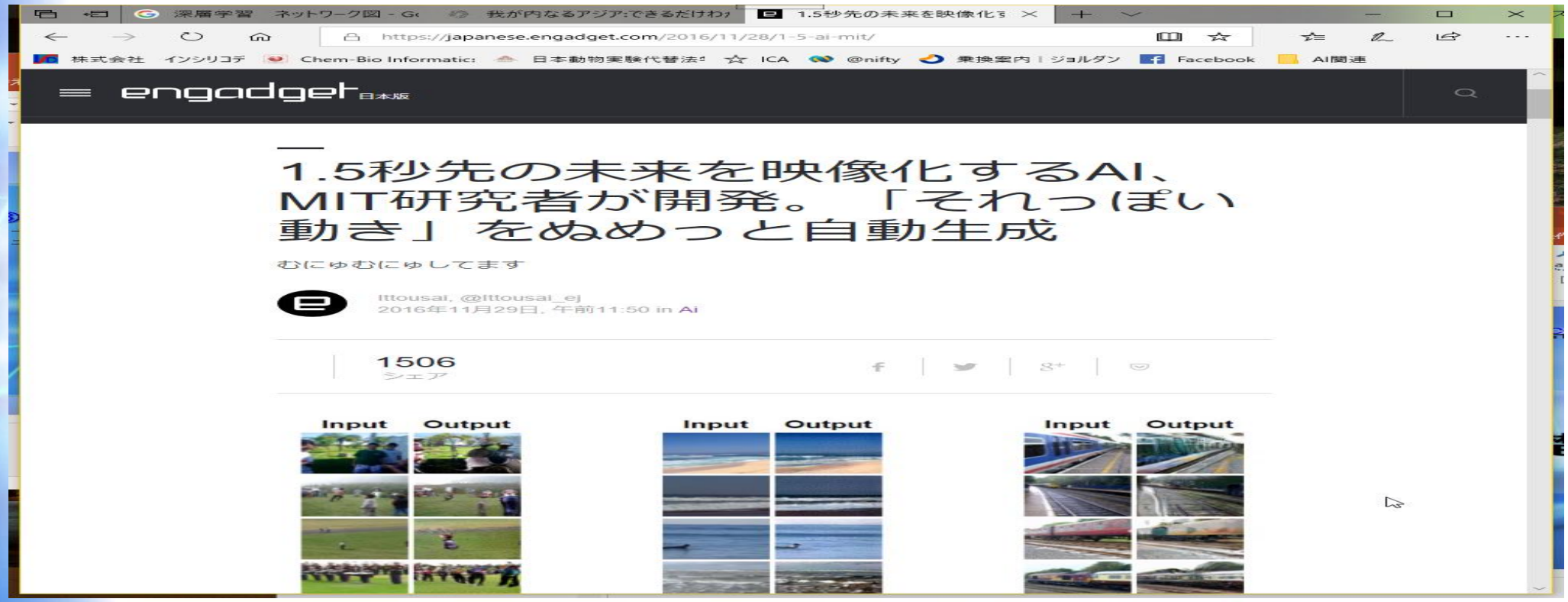
□化学分野で人工知能を適用する時の注意とまとめ

・機械学習型人工知能の有する基本的弱点

・学習した事や獲得情報以外への適用困難

ー秒後の状態認識できない⇒動く自動車の写真解析でー秒後を予測できない⇒動くものと動かないものを認識する学習必要

＊**サンプル数が十分**であっても、**学習で獲得できない**ものがある



https://japanese.engadget.com/2016/11/28/1-5-ai-mit/

□化学分野で人工知能を適用する時の注意とまとめ

・機械学習型人工知能の有する基本的弱点

◇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**

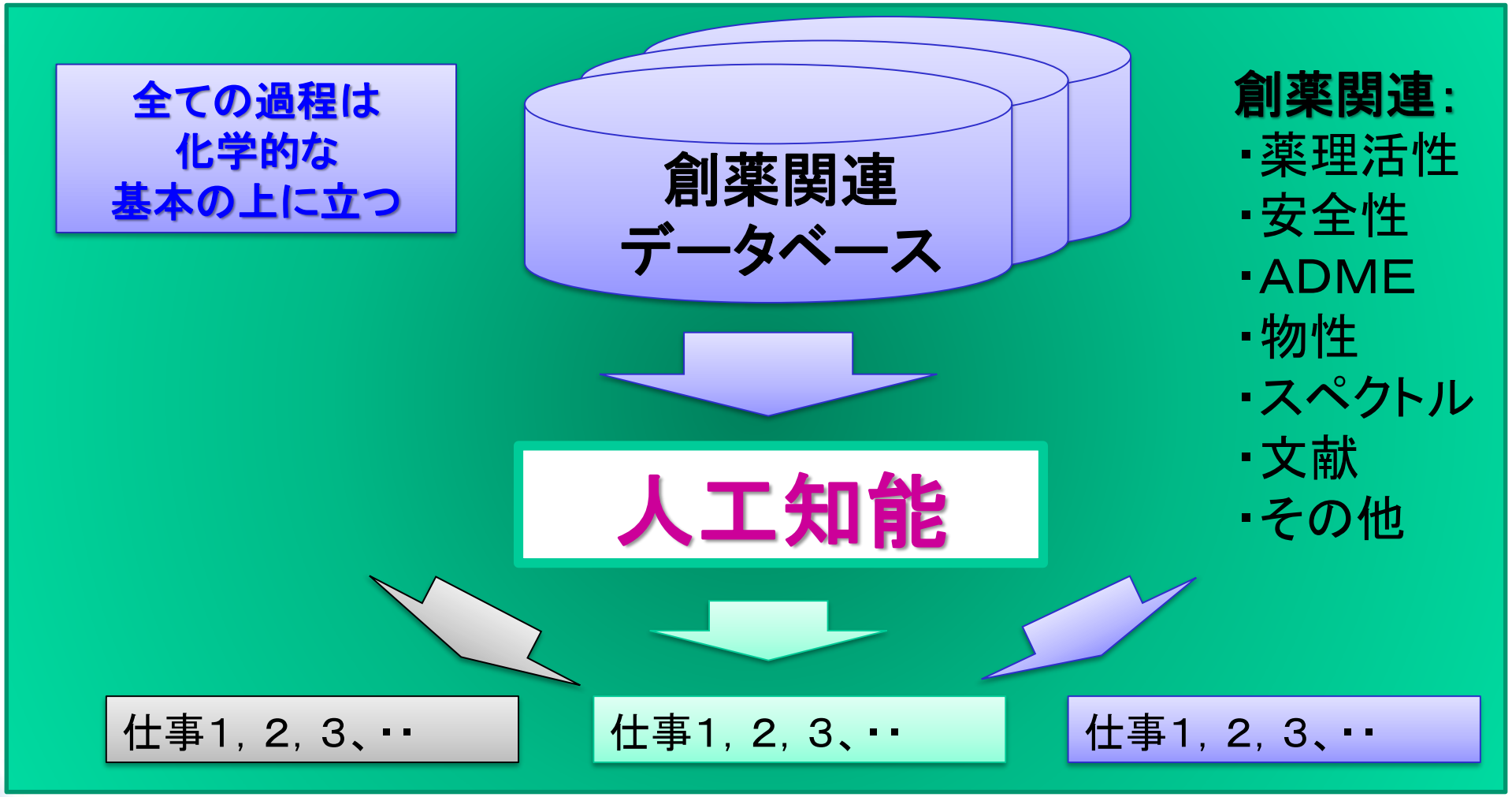
・構造－活性/毒性/物性相関等の研究では、要因解析が極めて大事である

◇**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

□化学分野で人工知能を適用する時の注意とまとめ

・人工知能適用における化学的問題



化学関連研究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

□化学分野で人工知能を適用する時の注意とまとめ

・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

・**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

□化学分野で人工知能を適用する時の注意とまとめ

・人工知能適用における化学的問題

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

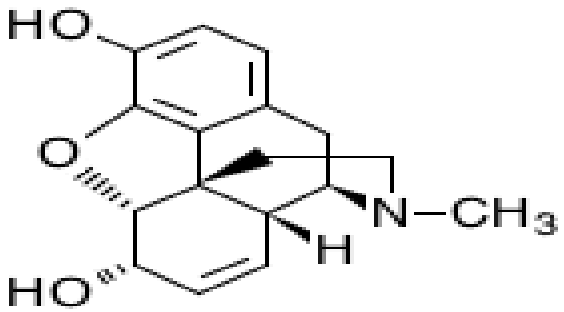
・化学やバイオ関連分野の論文は基本的に成功事例
　成功のみ掲載されている。このような成功事例のみを
　学習した結果提案される化合物は、
　成功／失敗化合物？　→　失敗というフィルターがない
・入力Journal数は精度の保証にならない
　数が多いほど上記の偏向学習が進んでいることの証拠

□化学分野で人工知能を適用する時の注意とまとめ

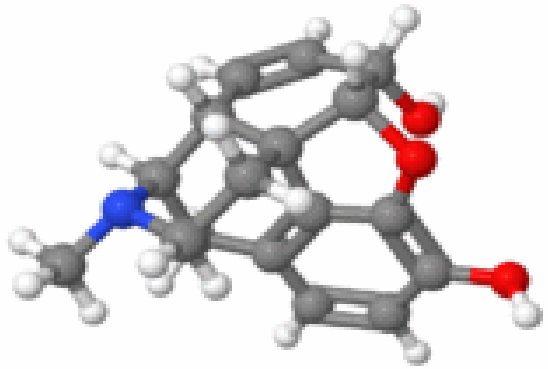
・人工知能適用における化学的問題

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：

```
  -ISIS-  0  0  1  1  0  0  0  0  0  0999 V2000
    21.5533  -23.4873    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
    21.5533  -22.2201    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
    22.7502  -24.1914    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
    20.4268  -24.1210    0.0000 C   0  0  1  0  0  0  0  0  0  0  0  0
    22.7502  -22.8537    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    20.4268  -21.5864    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    22.7502  -21.5864    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    23.8062  -23.4873    0.0000 C   0  0  1  0  0  0  0  0  0  0  0  0
    22.7502  -25.4587    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    19.0891  -22.7833    0.0000 O   0  0  2  0  0  0  0  0  0  0  0  0
    20.4268  -25.4587    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
    24.9327  -22.8537    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    20.4268  -20.2487    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    23.8062  -22.2905    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    22.7502  -20.2487    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    24.9327  -24.1914    0.0000 N   0  0  3  0  0  0  0  0  0  0  0  0
    21.5533  -26.1627    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    19.3003  -26.1627    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    21.6237  -19.6151    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    19.3003  -19.6151    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    23.9470  -25.1067    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    26.3408  -24.1914    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0     0  0
  1  3  1  0     0  0
  1  4  1  0     0  0
  1  5  1  1     0  0
  2  6  2  0     0  0
  2  7  1  0     0  0
  3  8  1  0     0  0
  3  9  1  0     0  0
  4 10  1  6     0  0
  4 11  1  0     0  0
  5 12  1  0     0  0
  6 13  1  0     0  0
  7 14  1  0     0  0
  7 15  2  0     0  0
  8 16  1  1     0  0
  9 17  2  0     0  0
 11 18  1  6     0  0
 13 19  2  0     0  0
 13 20  1  0     0  0
  6 10  1  0     0  0
  8 14  1  0     0  0
 11 17  1  0     0  0
 12 16  1  0     0  0
 16 19  1  0     0  0
  3 21  1  1     0  0
 16 22  1  0     0  0
M  END
```

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

□化学分野で人工知能を適用する時の注意とまとめ

・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**

□化学分野で人工知能を適用する時の注意とまとめ

・機械学習における問題

□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

□深層学習実施上での留意点

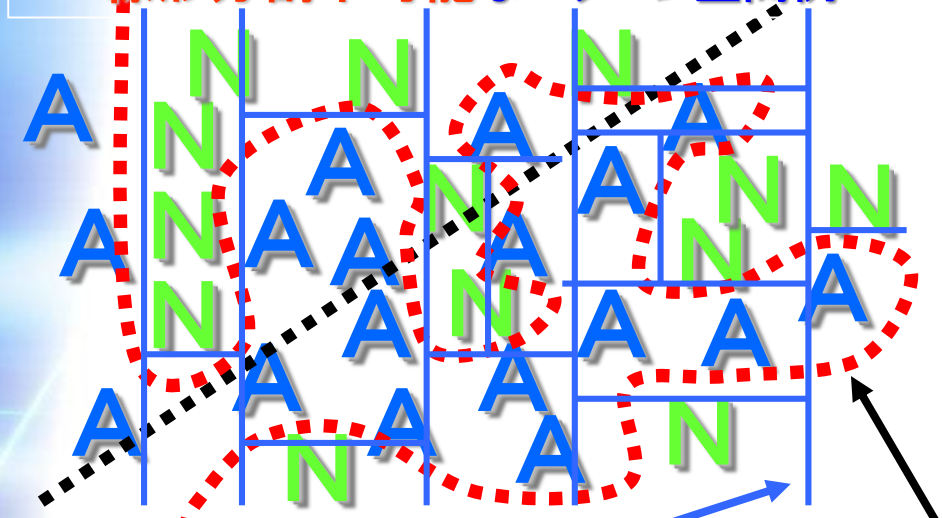
1. 学習に用いるサンプル数問題→過剰適合の回避
   ネットワークの階層が深いと、学習に必要なサンプル数が急激に増大
   現時点（2016.11.13）でのCAS登録化合物（有機／無機）数
   ⇒123,623,093
2. 学習の偏り回避→サンプルの学習内容が大事
3. 現在は画像／音声／文字認識が主体のネットワーク構成
4. 結果が良くても、ネットワークから要因情報を取り出すことが
   極めて困難

□化学分野で人工知能を適用する時の注意とまとめ

- 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

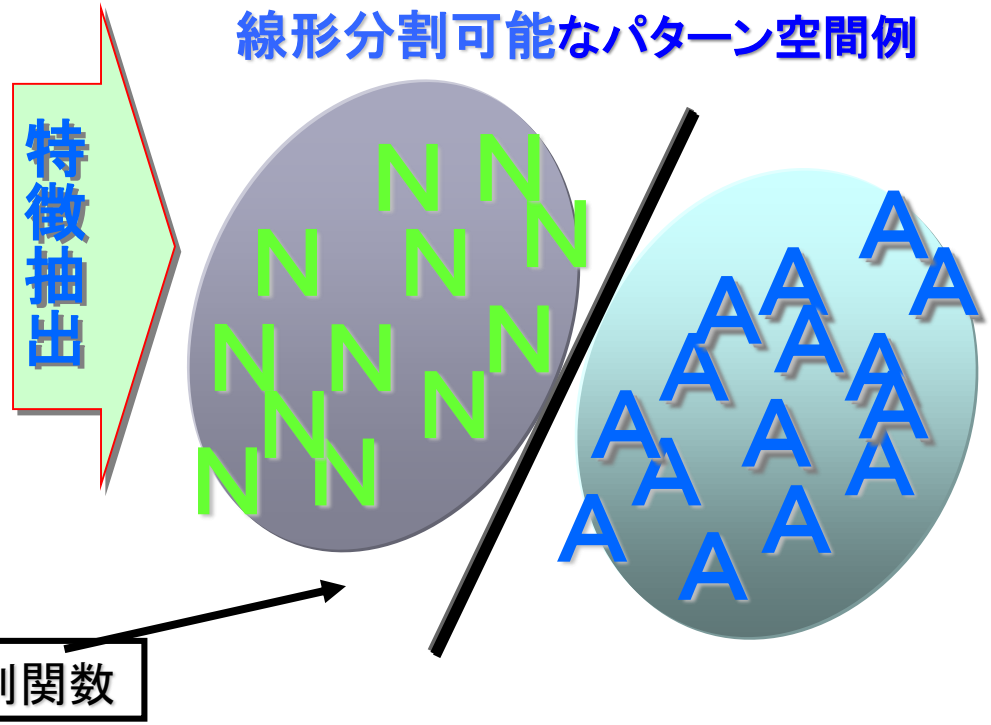
線形分割不可能なパターン空間例



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

線形分割可能なパターン空間例



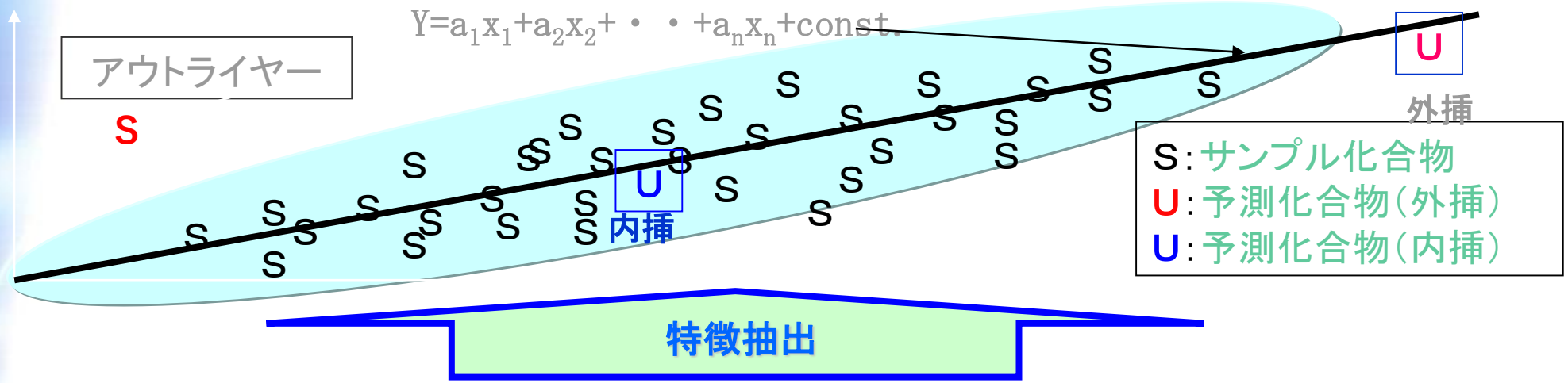
分類目的を達成する空間構築による分類

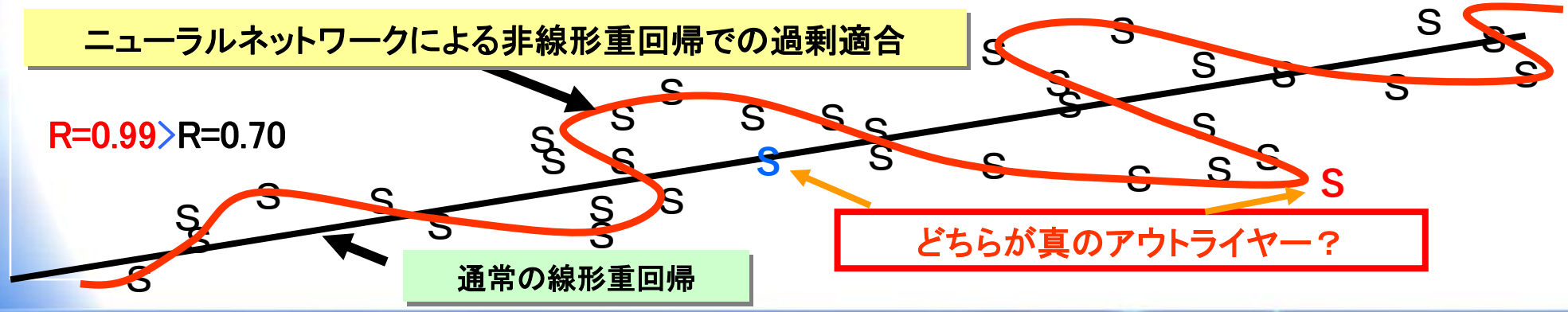
空間にAとNを分ける情報を持つ
空間を形成するパラメータ群は
AとNの識別根拠情報が見えやすい

□化学分野で人工知能を適用する時の注意とまとめ

・機械学習における問題

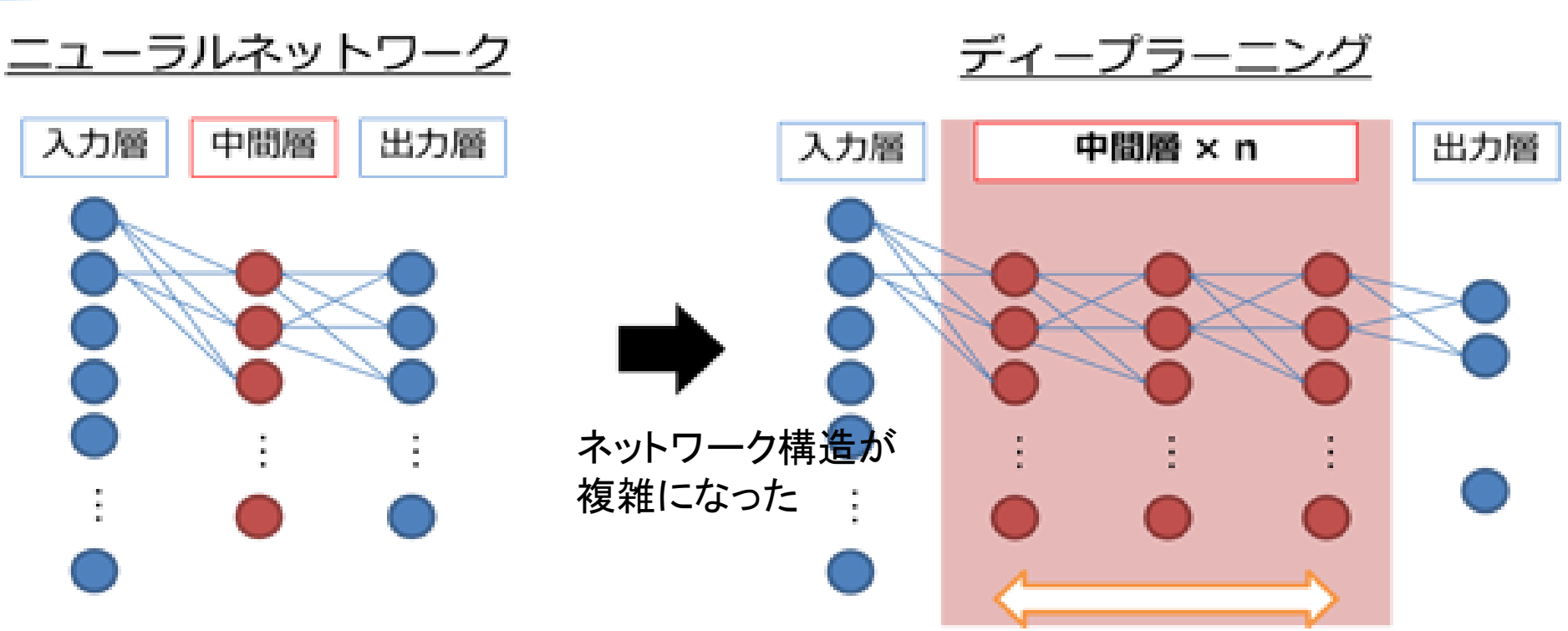
線形重回帰における過剰適合問題





□化学分野で人工知能を適用する時の注意とまとめ

・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

□化学分野で人工知能を適用する時の注意とまとめ

・機械学習型人工知能適用上での解決すべき点

In Silico Data
Miracles by the KY-methods

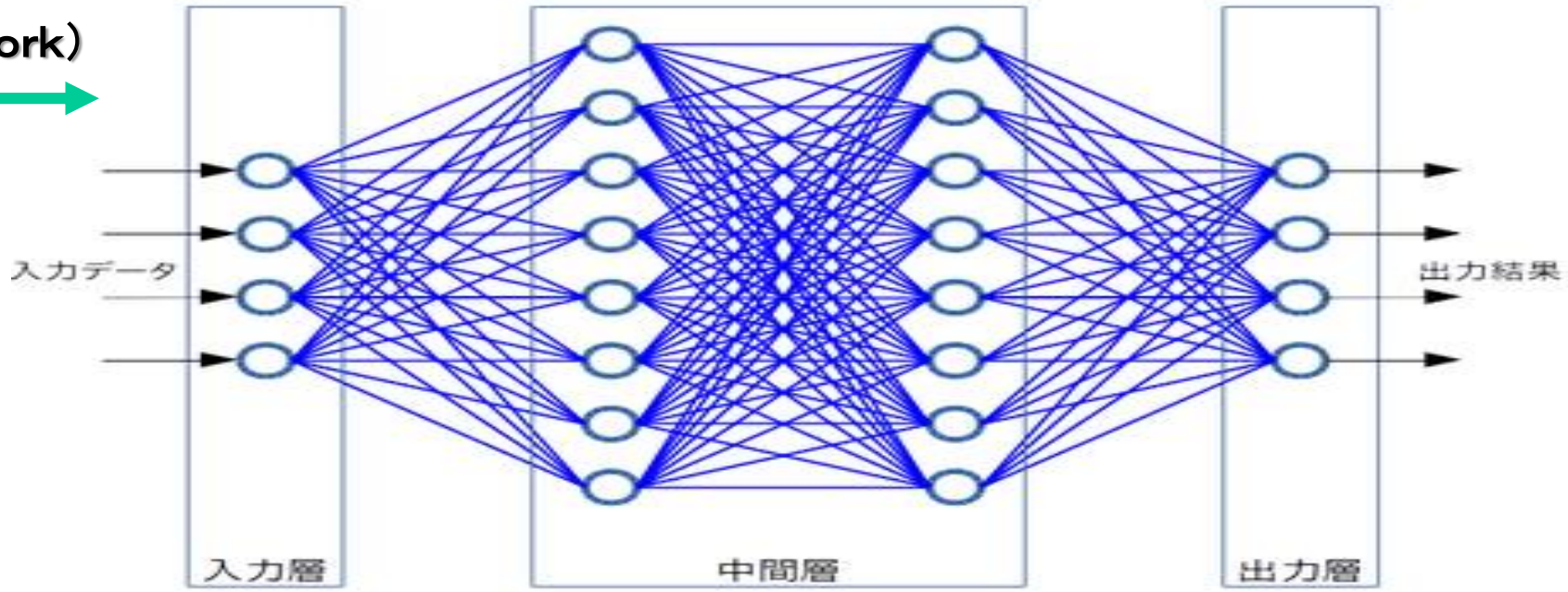
④ネットワーク構造が極めて複雑なので、要因解析ができない

・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

＊ニューラルネットワーク発表当時、構造－活性相関分野では
この説明困難な事実が、大きな問題となった

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

□化学分野で人工知能を適用する時の注意とまとめ
・生命科学分野における今後の人工知能の展開

□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
□Iotや分析／医療機器との連携

□その他の展開

□化学分野で人工知能を適用する時の注意とまとめ
・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

□　機械学習（深層学習）実施上での問題点

①サンプル数と量

ニューラルネットワークである限り、膨大な数のサンプルが必要

②サンプルの品質と内容

サンプルは解析目的に従った情報を有する必要がある

サンプルはデータの偏りのないことが望ましい

＊ネガデータサンプルだけの情報で、ポジを検討できない

＊間違った情報を学習させない⇒ヘイト発言をするロボット

③学習した事や獲得情報以外への適用困難

一秒後の状態認識できない⇒動く自動車の写真解析で一秒後を予測できない⇒動くものと動かないものを認識する学習必要

④ネットワーク構造が複雑なので、要因解析ができない

＊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

□化学分野で人工知能を適用する時の注意とまとめ

・AIを正しく適用し信頼性の高い結果を得るために

In Silico Data
Miracles by the KY-methods

□　機械学習（深層学習）実施上での問題点

①サンプル数と量

ニューラルネットワークである限り、膨大な数のサンプルが必要

②サンプルの品質と内容

サンプルは解析目的に従った情報を有する必要がある

サンプルはデータの偏りのないことが望ましい

＊ネガデータサンプルだけの情報で、ポジを検討できない

＊間違った情報を学習させない⇒ヘイト発言をするロボット

③学習した事や獲得情報以外への適用困難

一秒後の状態認識できない⇒動く自動車の写真解析で一秒後を予測できない⇒動くものと動かないものを認識する学習必要

④ネットワーク構造が複雑なので、要因解析ができない

＊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

In Silico Data
Miracles by the KY-methods

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