◇第三部　人工知能（Artificial Intelligence）関連
Part3. Artificial Intelligence

株式会社　インシリコデータ
湯田　浩太郎

Contents:

生命科学分野での人工知能への期待分野と適用／成果
Expected fields and applications / results for artificial intelligence in the life science field

	化学／創薬 Chemistry / Drug design	バイオ Bio	医療 Medical care
適用分野 Application fields			
人工知能 Artificial intelligence	初期Initial⇒	ルールベース型 Rule base type	
	現在Current⇒	機械学習型（深層学習）Machine deep learning	
期待成果 Expected results	新薬デザイン、最適化、ドラグ・リポジショニング、毒性評価、New drug design, optimization, drag repositioning, toxicity assessment, etc.	遺伝子解析、遺伝子探索、発現プロフィール解析、SNPs探索、Gene analysis, gene search, expression profile analysis, SNPs search, etc.	自動診断、画像解析、音声解析、その他 Automatic diagnosis, image analysis, audio analysis, etc.

□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 Rule-based artificial intelligence

DEREK ⇒ 化合物の毒性評価システム Compound toxicity evaluation system

☆コグニティブコンピュータ Cognitive computer

Watson (IBM) ⇒クイズ番組で勝利 Win a quiz show
医療関連分野で実績を出しつつあるProven in the medical field

☆機械学習（深層学習）型システム Machine/ Deep learning

アルファ碁 AlphaGo（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？

Win 4 wins and 1 loss against the world's top professional athletes
・Number of learning: tens of millions of stations> 30 million stations (self-playing)
・Automatic rule acquisition: Did you automatically learn Go rules?

☆機械学習（深層学習）型システムMachine/ Deep learning

アルファ碁 AlfaGo（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
Win 4 wins and 1 loss against the world's top professional athletes

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
The cause of victory? : Changed from a rule type game to a kind of pattern recognition type game
As a result, it overturned the evaluation that it would take 10 years to win the game.

□対局の条件と特徴的結果　Game conditions and characteristic results

・学習サンプル数：数千万局＞三千万局（自己対局）

Number of learning samples: tens of millions> 30 million (self-playing)

・ルールの自動獲得：囲碁のルールを自動的に学習した？

Automatic rule acquisition: Did you automatically learn Go rules?

上記事実から受ける期待イメージ：　Expected image from the above facts

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？

If deep learning is performed using a large amount of data, can some new and important rules be discovered, leading to new research?

□　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

■ 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

□人工知能概論と化学分野の人工知能

Introduction to artificial intelligence and artificial intelligence in the chemical field

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

□人工知能概論と化学分野の人工知能

Introduction to artificial intelligence and artificial intelligence in the chemical field

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

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

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:

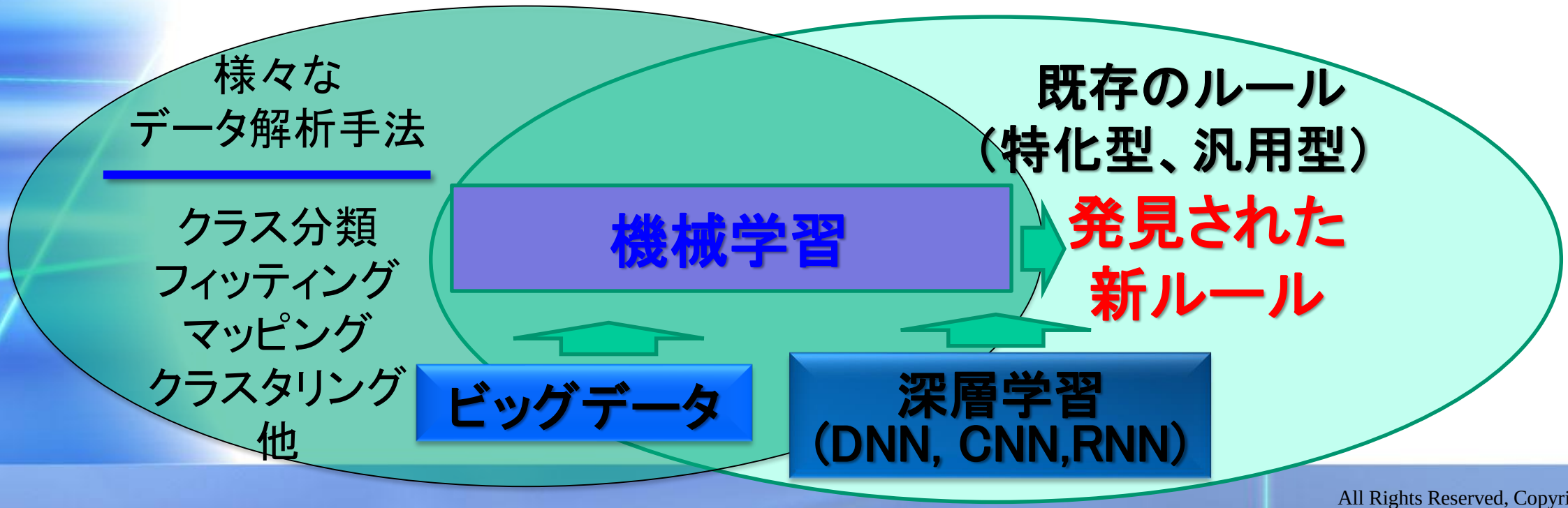
RIPT: **R**ule **I**nduction for **P**redictive **T**oxicology

TOX-MATCH:

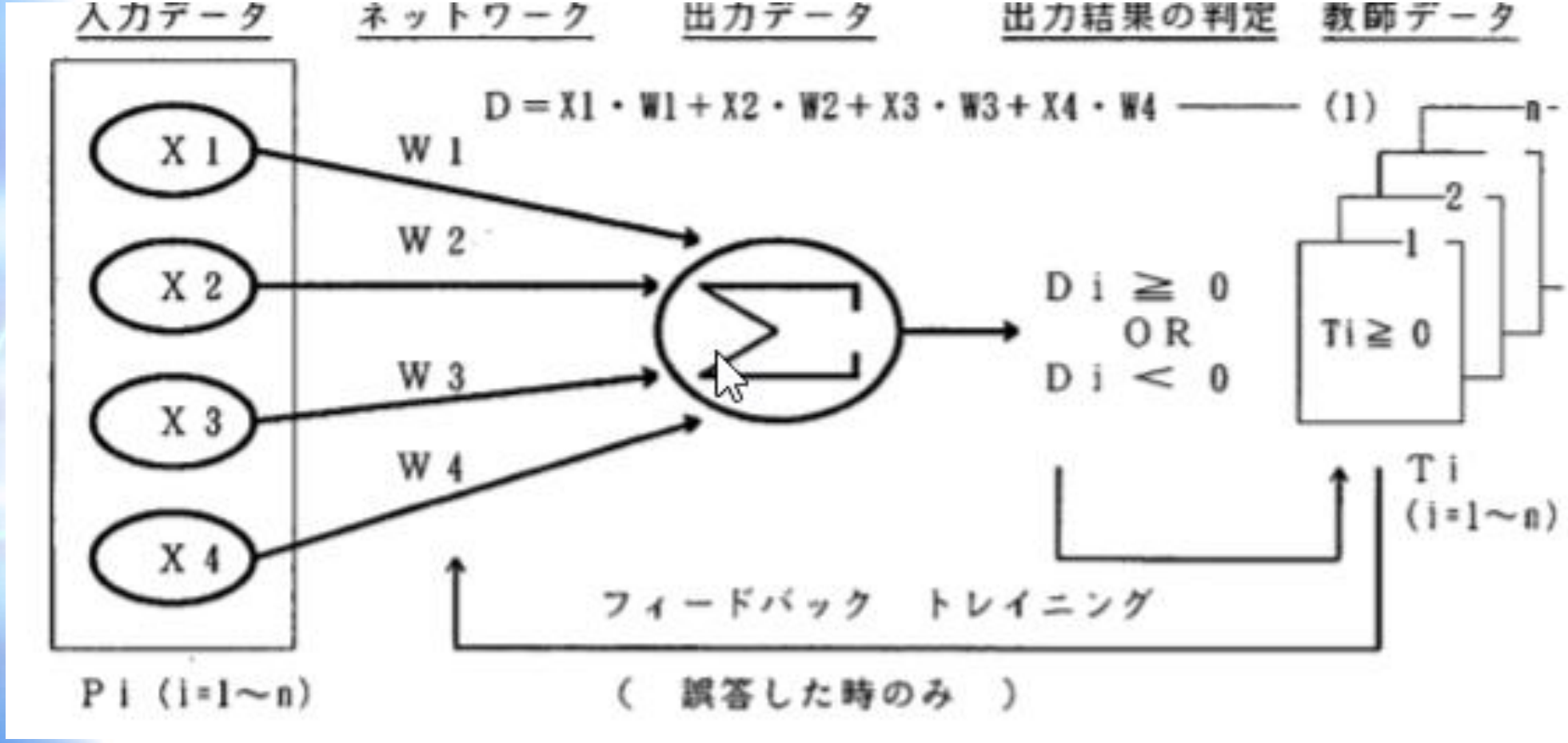
DART: **D**ecision **A**nalysis by **R**anking **T**echniques

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

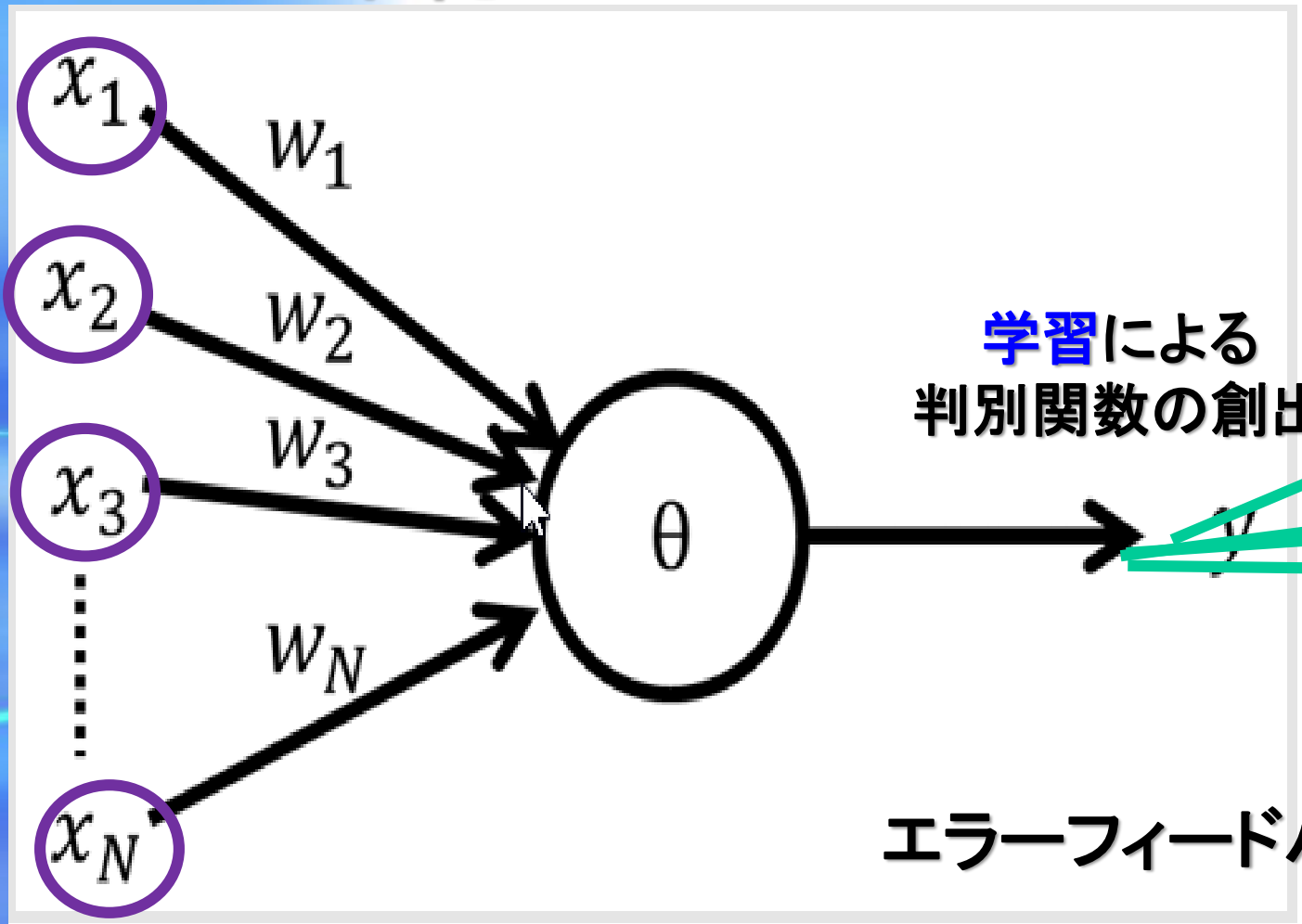
多変量解析/パターン認識と人工知能は
機械学習により繋がっている



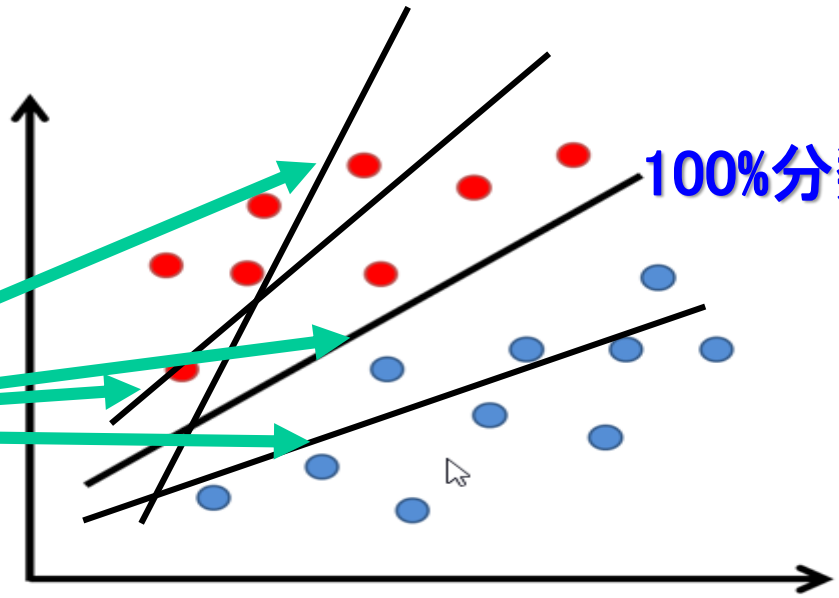
パーセプトロンの学習アルゴリズム



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

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

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

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

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

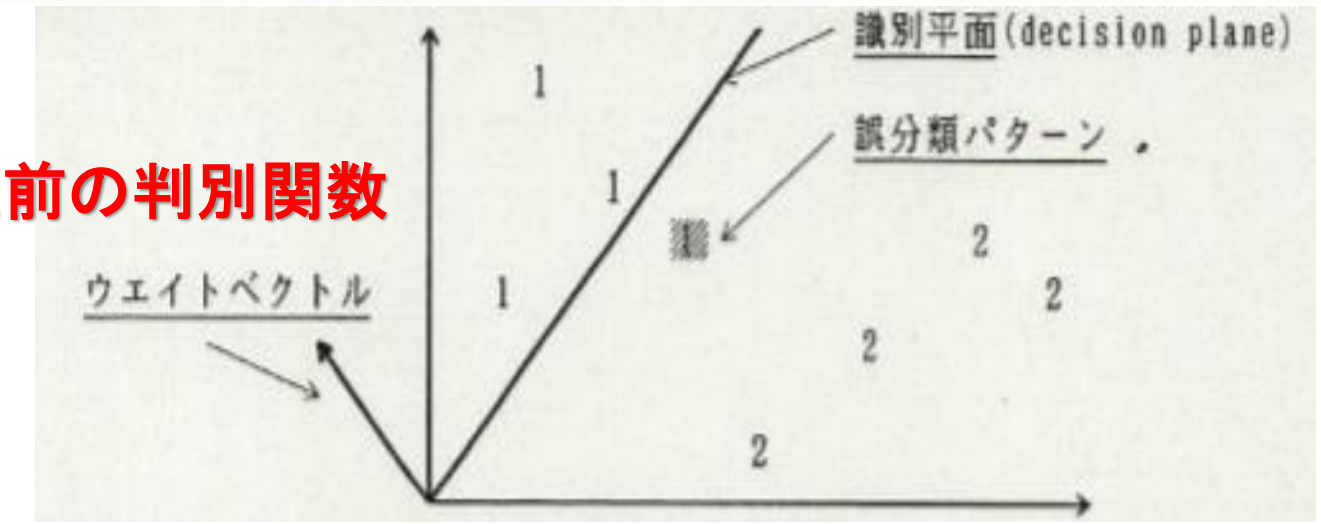
、（5）式のWnを（4）式に代入し、Cについて変換すると、

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

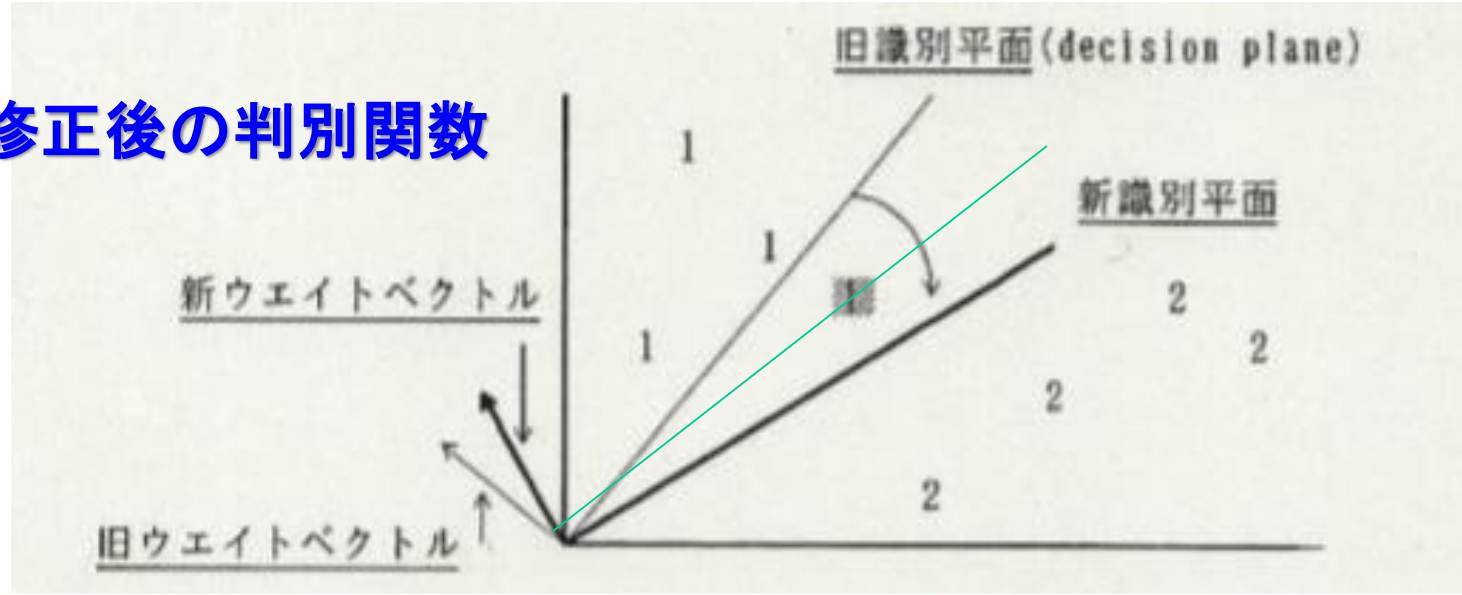
と（4）式から $D' = -Wo \cdot Xp$ 従って、

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

修正前の判別関数



修正後の判別関数



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。

バックプロパゲーション計算式の一部

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、

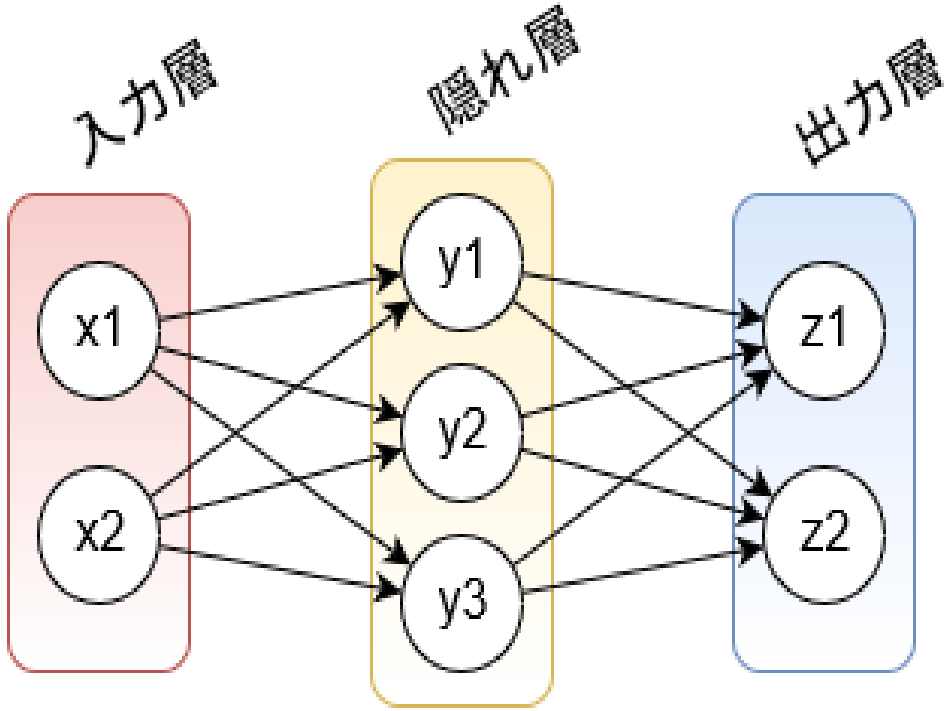
$$\frac{\partial L}{\partial b_{z_2}} = \frac{\partial L}{\partial z_2} \cdot \frac{\partial z_2}{\partial I_{z_2}} \cdot \frac{\partial I_{z_2}}{\partial b_{z_2}} = \frac{\partial L}{\partial z_2} \cdot \sigma_z'(I_{z_2}) \quad \left(\because \frac{\partial I_{z_2}}{\partial b_{z_2}} = 1 \right) \tag{16}$$

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

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

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

となります。



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

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　　深層学習

エラーフィードバックトレーニング、他

Bayes、SVM、AdaBoost、他

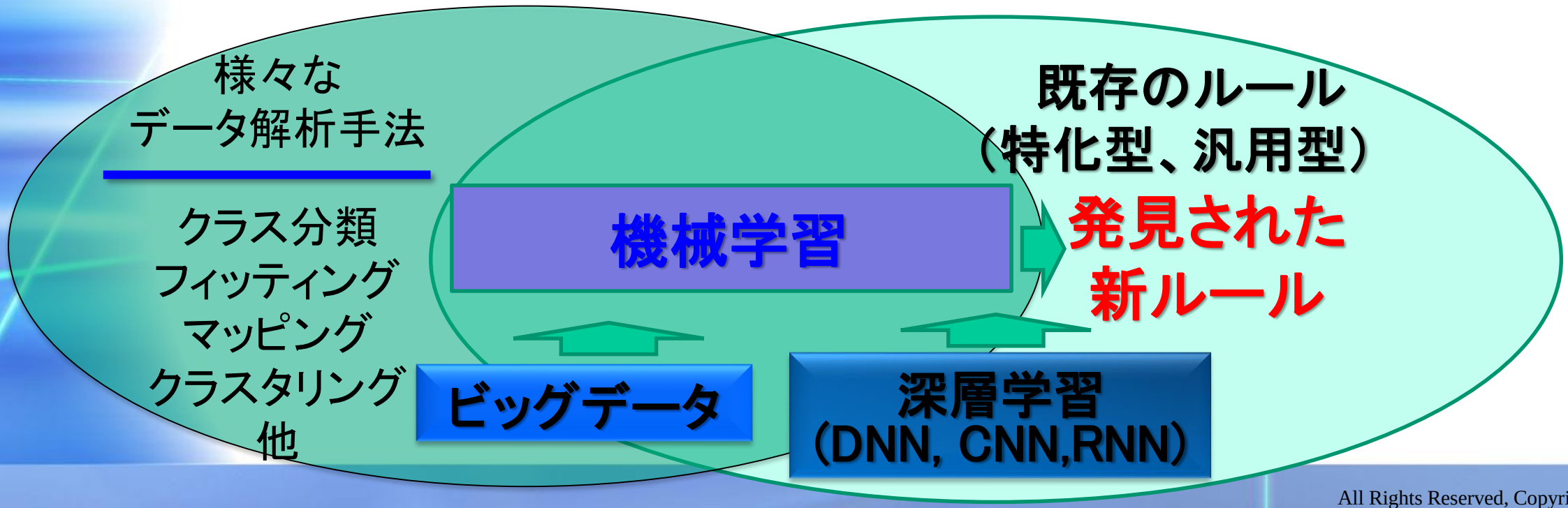
1. 人工知能概論と化学分野の人工知能

・ケモメトリックスと人工知能の関係

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

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

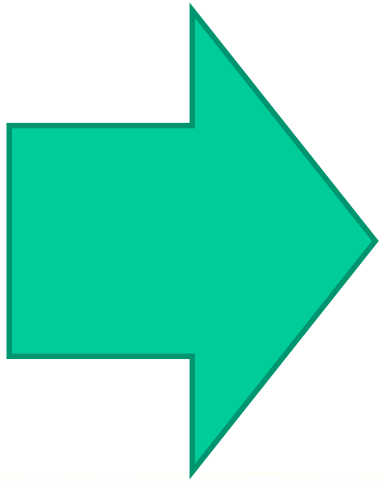
・特徴：人間のノウハウの活用

◇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

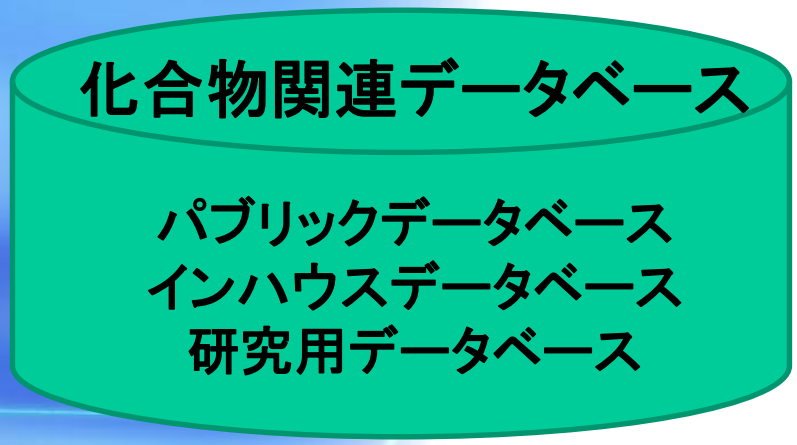
構造－毒性相関
化合物毒性評価/予測
脱毒性デザイン

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

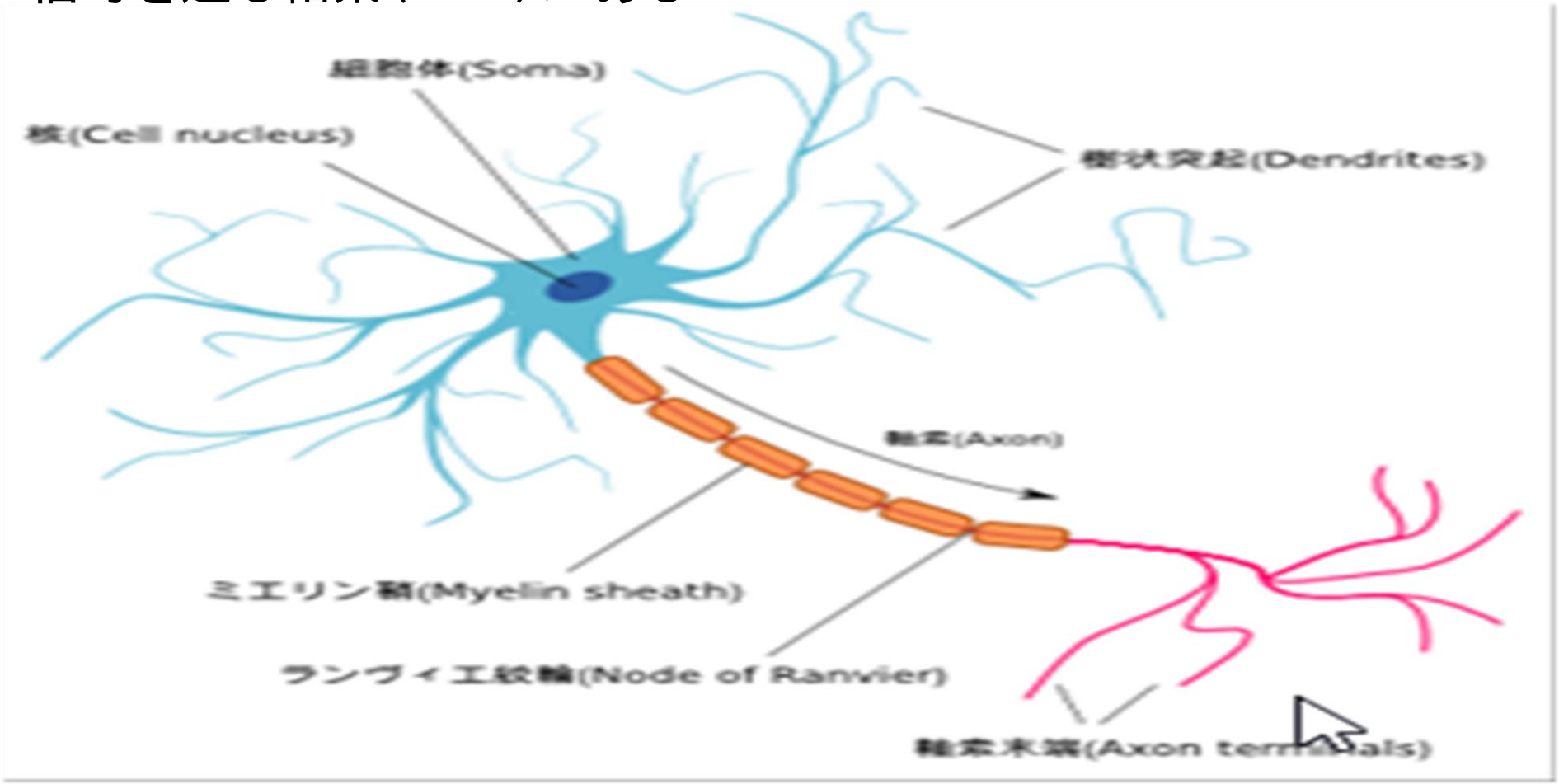
構造－毒性相関
化合物毒性評価/予測
脱毒性デザイン

構造－物性相関
機能性化合物デザイン

□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

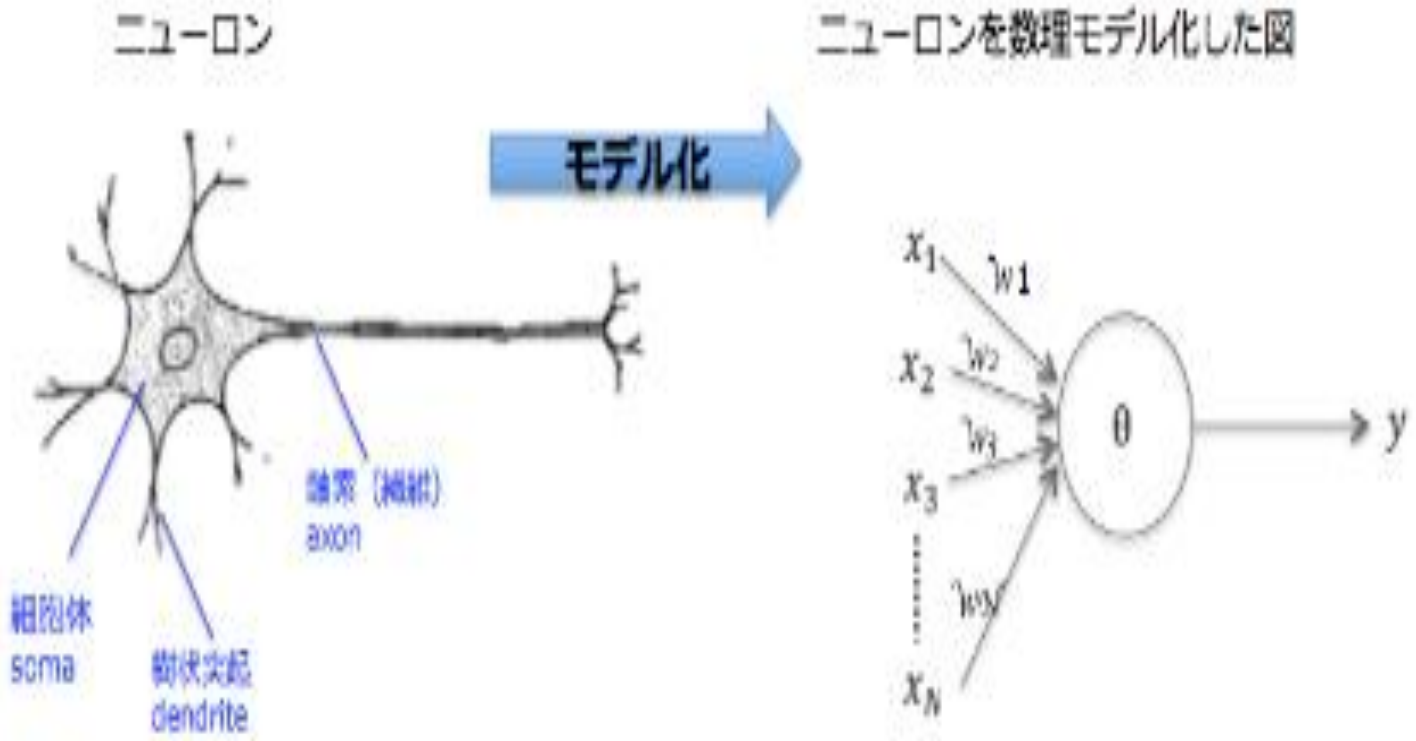


http://kazoo04.hatenablog.com/entry/agi-ac-3

□機械学習型人工知能 Machine learning type artificial intelligence

特徴：ニューロンを用いた脳のシミュレーション
Features: Brain simulation using neurons

パーセプトロン
perceptron



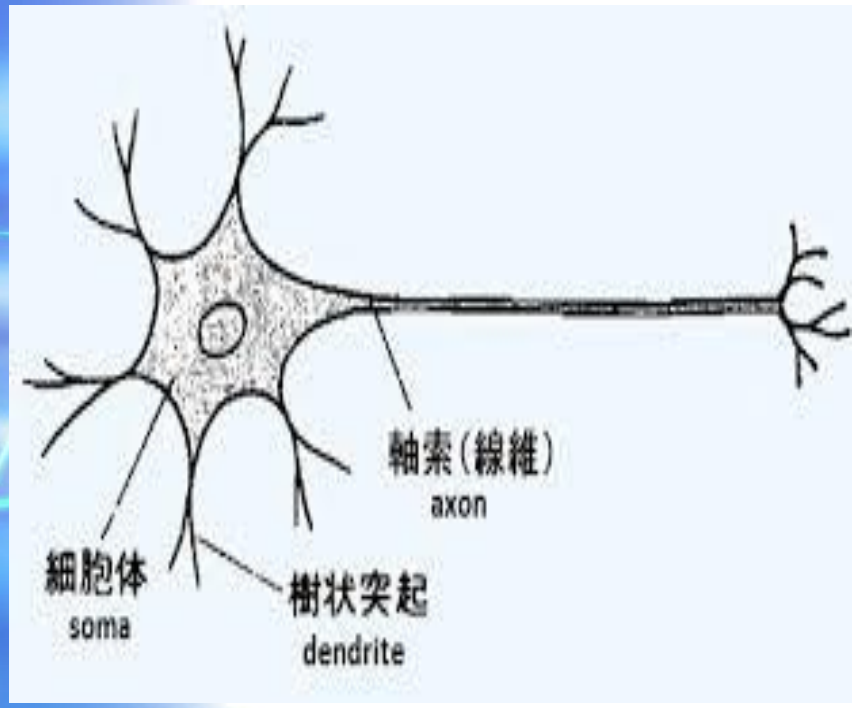
https://udemy.benesse.co.jp/ai/neural-network.html

□機械学習型人工知能 Machine learning type artificial intelligence

特徴：ニューロンを用いた脳のシミュレーション

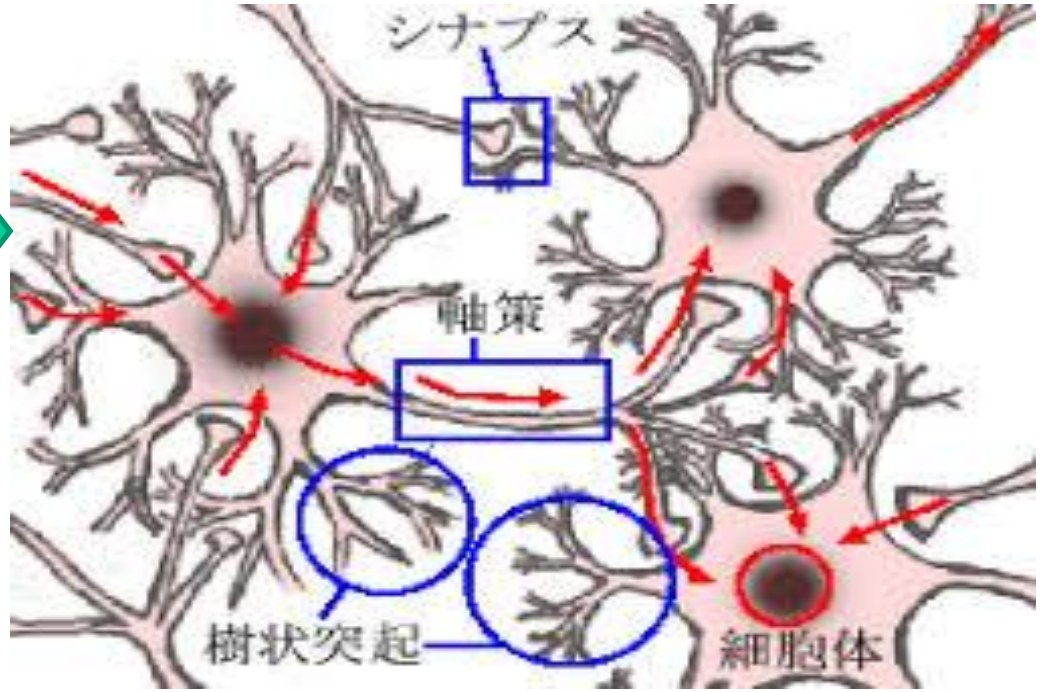
Features: Brain simulation using neurons

単ニューロンモデル
Single neuron model



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

ネットワークニューロンモデル
Network neuron model



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

□機械学習型人工知能 Machine learning type artificial intelligence

単ニューロンモデル
Single neuron model

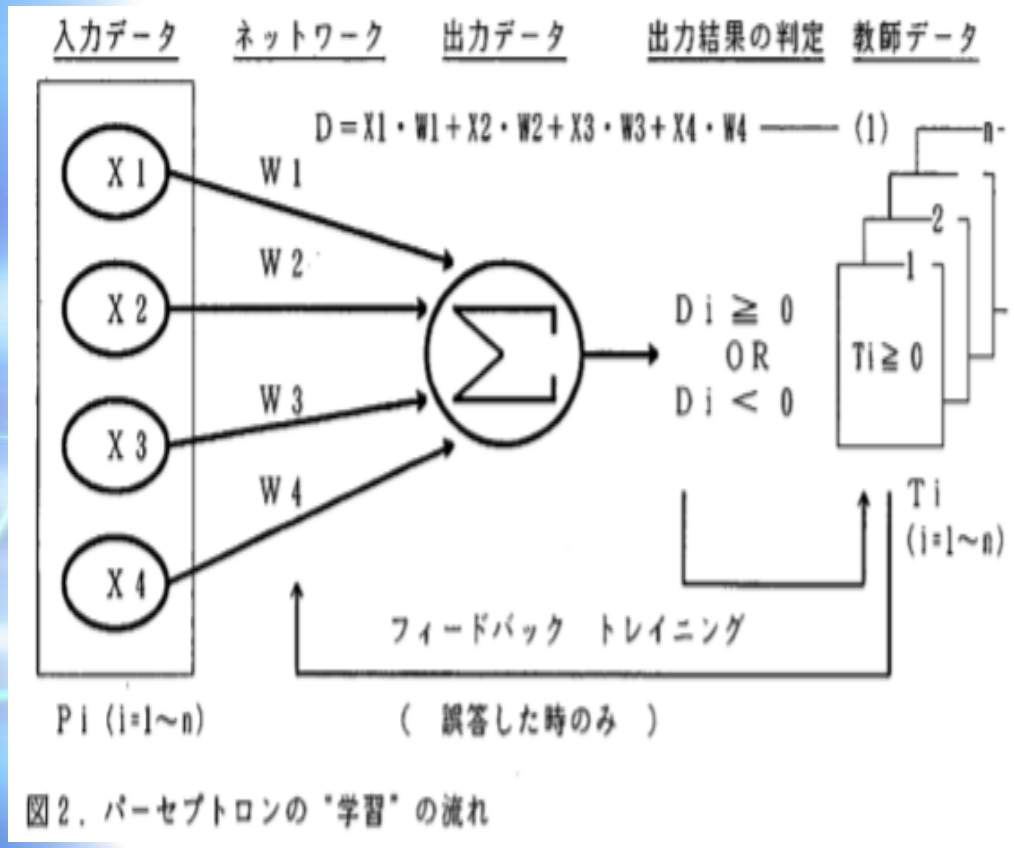
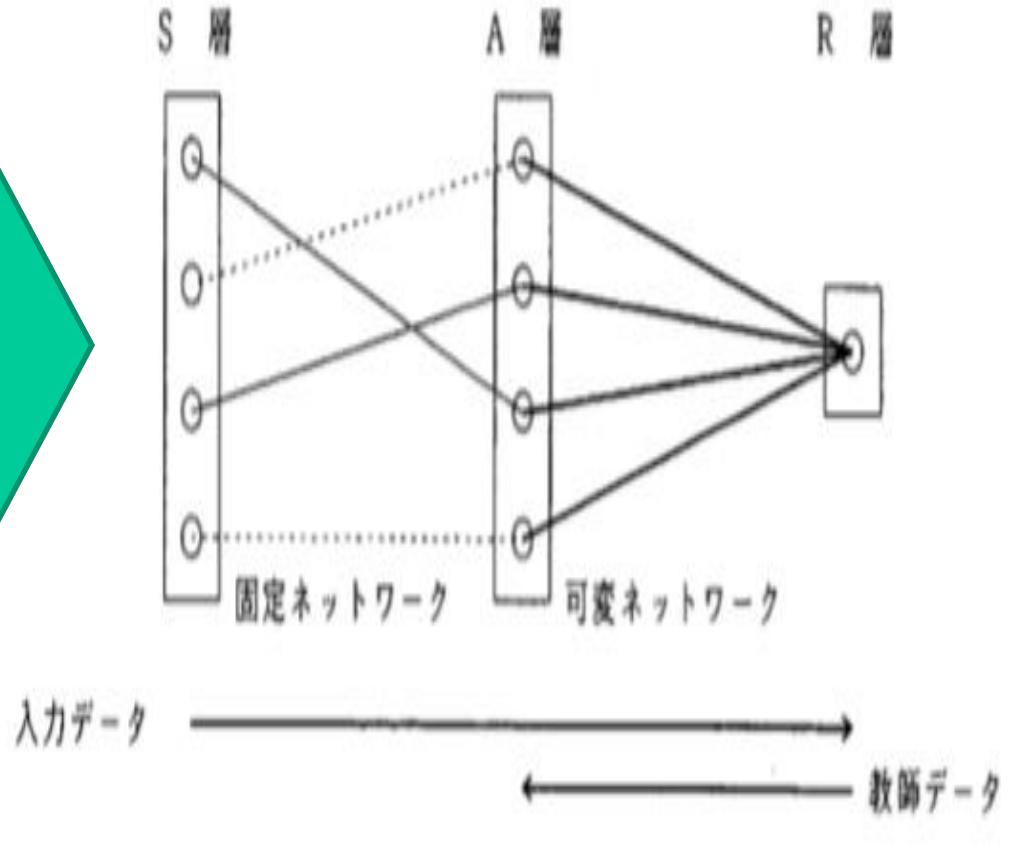


図2．パーセプトロンの"学習"の流れ

ネットワークニューロンモデル
Network neuron model



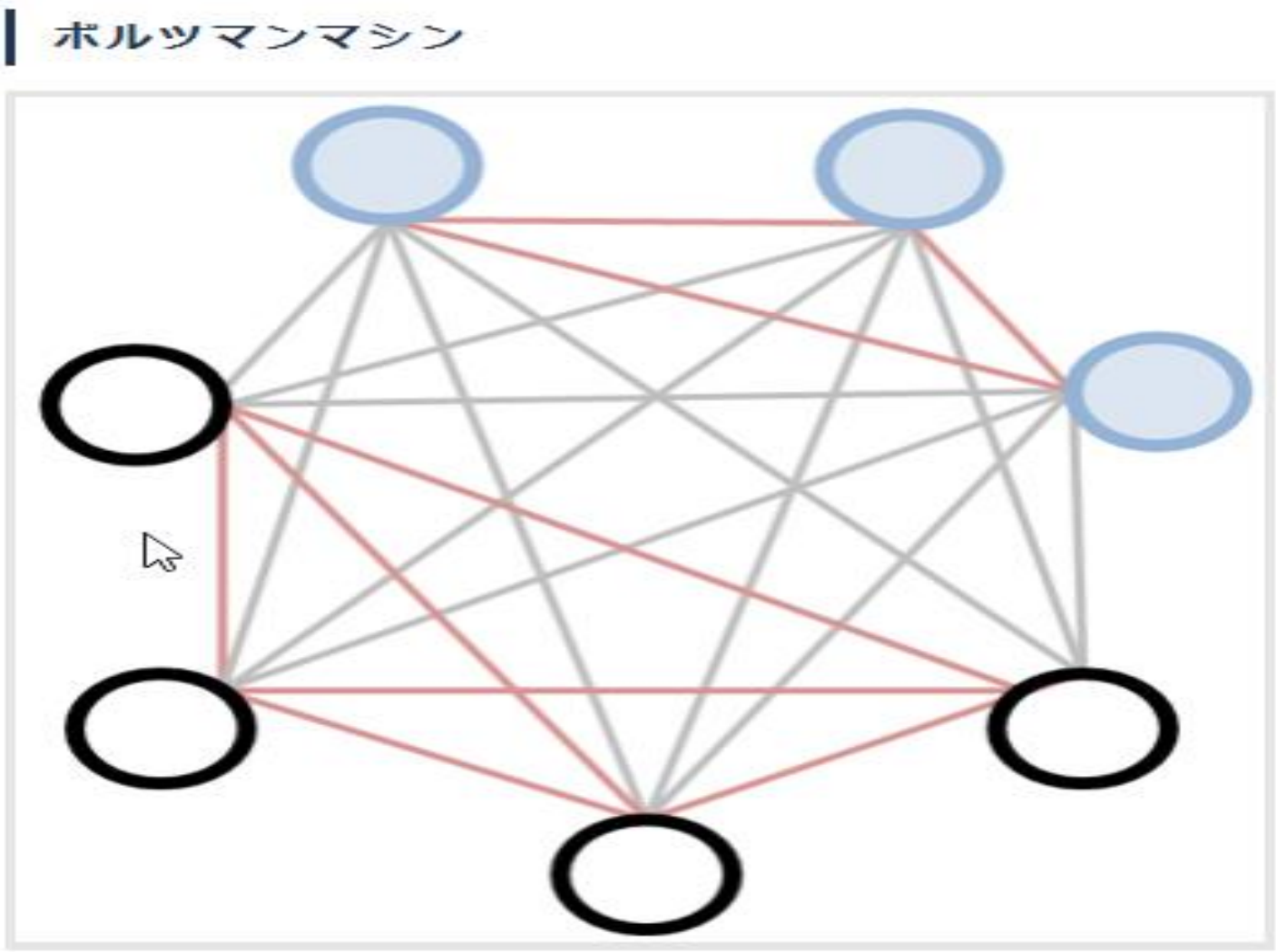
パーセプトロン Perceptron　　ニューラルネットワーク N.N.

□機械学習型人工知能 Machine learning type artificial intelligence

◇閉鎖型ニューラルネットワーク
Closed neural network

ボルツマンマシン
Boltzmann machine

最適化などに利用される
Used for optimization



パーセプトロンとニューラルネットワーク　Perceptron and neural network

＊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

＊Neural networks are based on a perceptron that uses a simple network structure.
＊Perceptron declines because it cannot solve simple two-classification problems
＊A neural network is proposed as an approach that breaks the limits of perceptron.

パーセプトロンとニューラルネットワーク　Perceptron and neural network

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

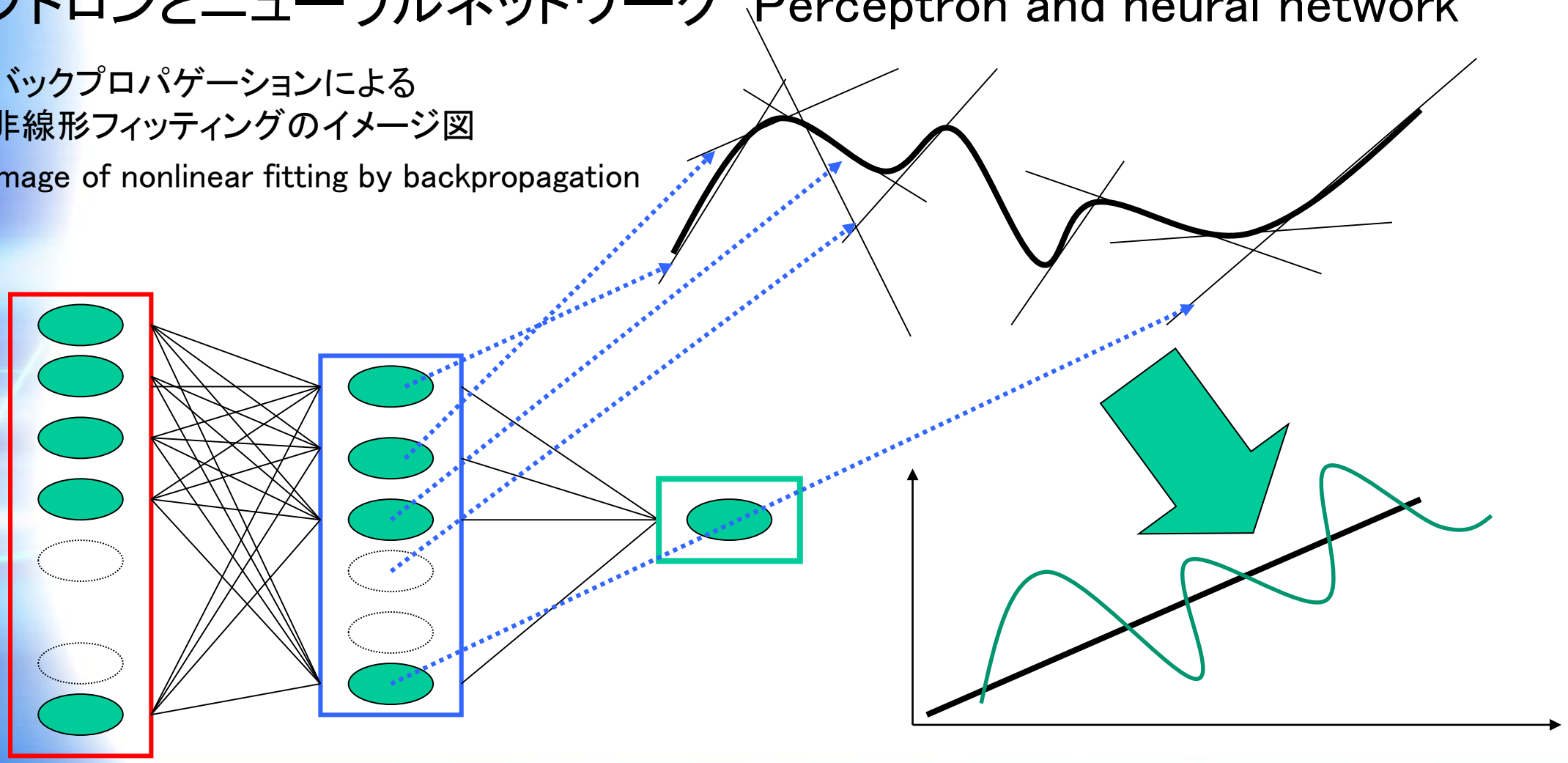
＊パーセプトロンは脳の機能を模したアプローチとして開発され、最終目標は
AI(人工知能）への展開であった
＊ニューラルネットワークはその改良系で、ネットワーク構造が多層構造となった
＊最近の深層学習はニューラルネットワークのネットワーク構造をさらに複雑にした

* Perceptron was developed as an approach that mimics the function of the brain, and the final goal was to develop AI (artificial intelligence)
* Neural network is an improved version of the network structure.
* Recent deep learning has further complicated the network structure of neural networks.

パーセプトロンとニューラルネットワーク Perceptron and neural network

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

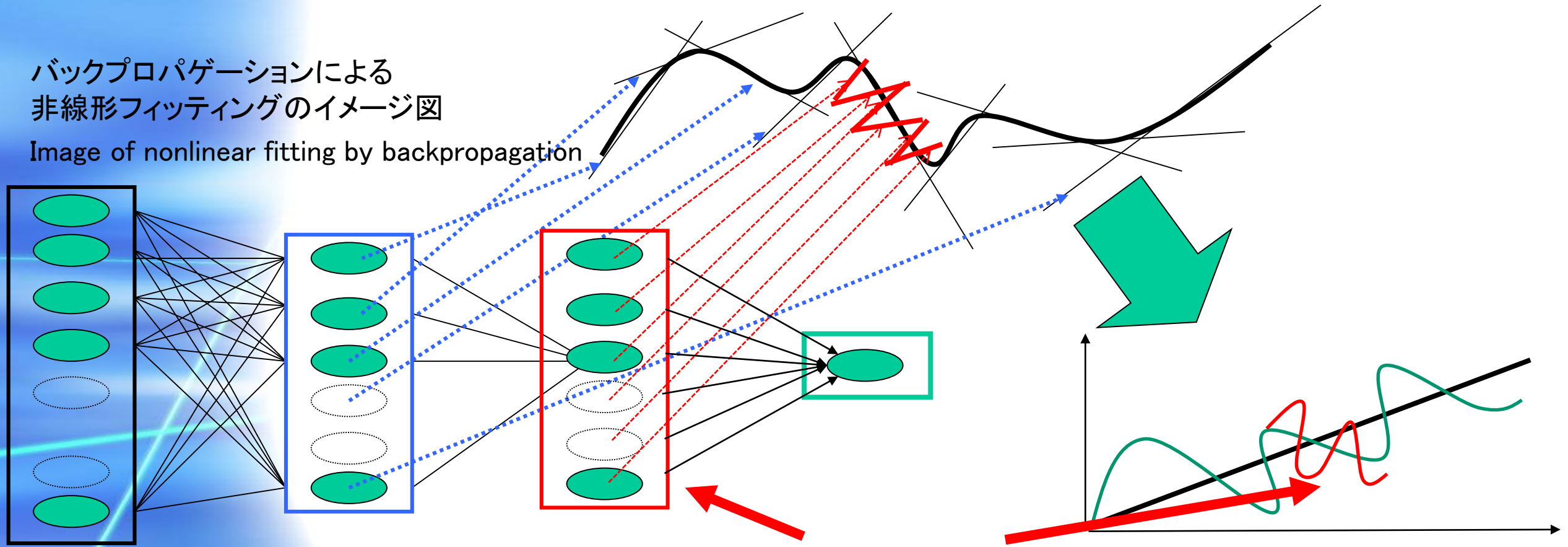
Image of nonlinear fitting by backpropagation



パーセプトロンとニューラルネットワーク Perceptron and neural network

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

Image of nonlinear fitting by backpropagation



中間層の段数が大きくなると非線形性が強くなる

・ニューラルネットワークのバックプロパゲーション Backpropagation of neural networks

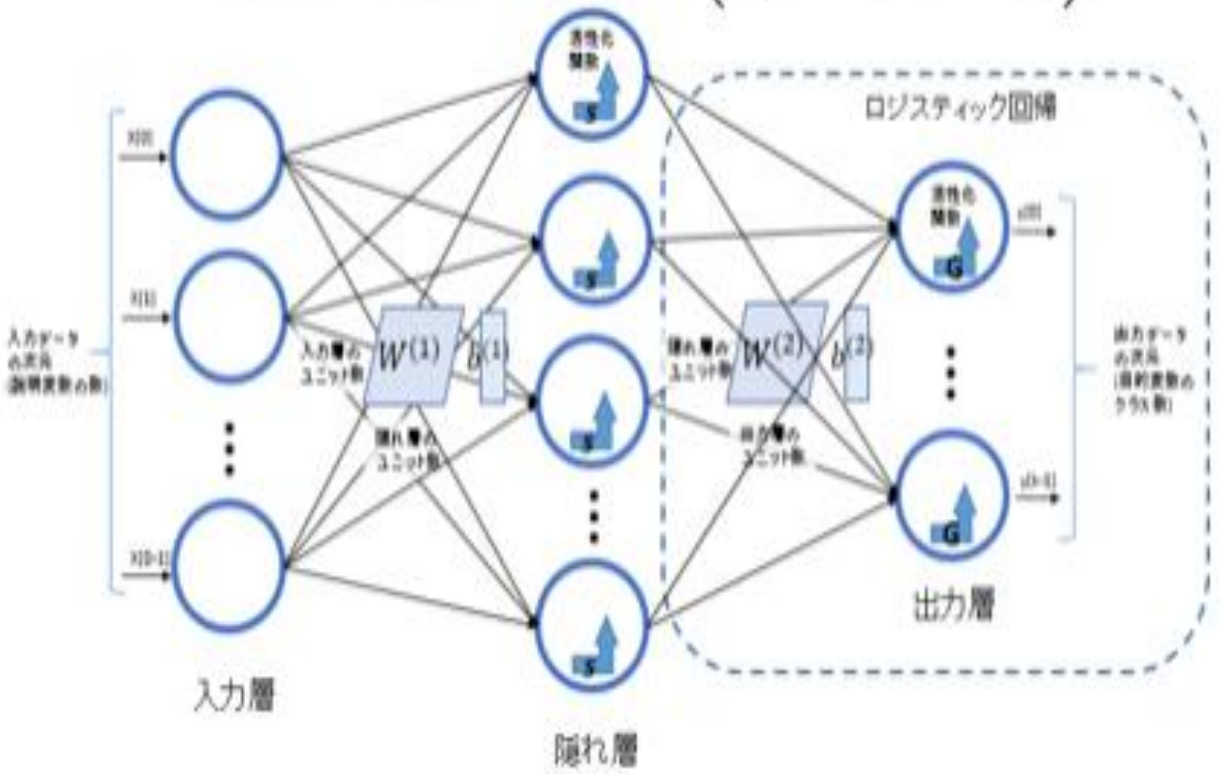
- $W^{(1)}$: 入力層 - 隠れ層の間で適用される係数行列。次元は 入力データの説明変数の数 D x 隠れ層のユニット数 D_h
- $b^{(1)}$: 入力層 - 隠れ層の間で適用される重みベクトル。次元は 隠れ層のユニット数 D_h
- s : 隠れ層の活性化関数。シグモイド関数もしくは $tanh$ (ハイパボリックタンジェント) をよく使う。
- $W^{(2)}$: 隠れ層 - 出力層の間で適用される係数行列 = ロジスティック回帰の係数行列。次元は 隠れ層のユニット数 x 出力データのクラス数
- $b^{(2)}$: 隠れ層 - 出力層の間で適用される重みベクトル =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

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



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

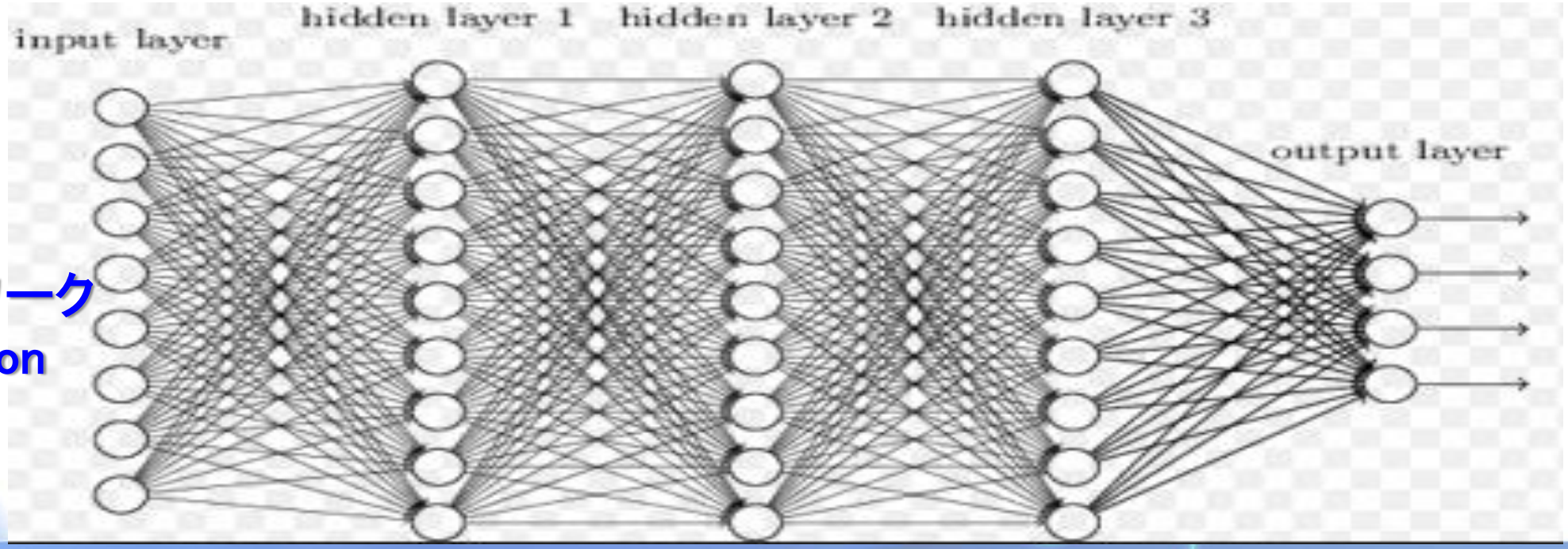
□機械学習型人工知能 Machine learning type artificial intelligence

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）
Execution environment: Development of machine learning methods (deep learning: deep learning)

◇ニューラルネットワークから深層学習へ　From neural network to deep learning

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
Forward propagation
neural network

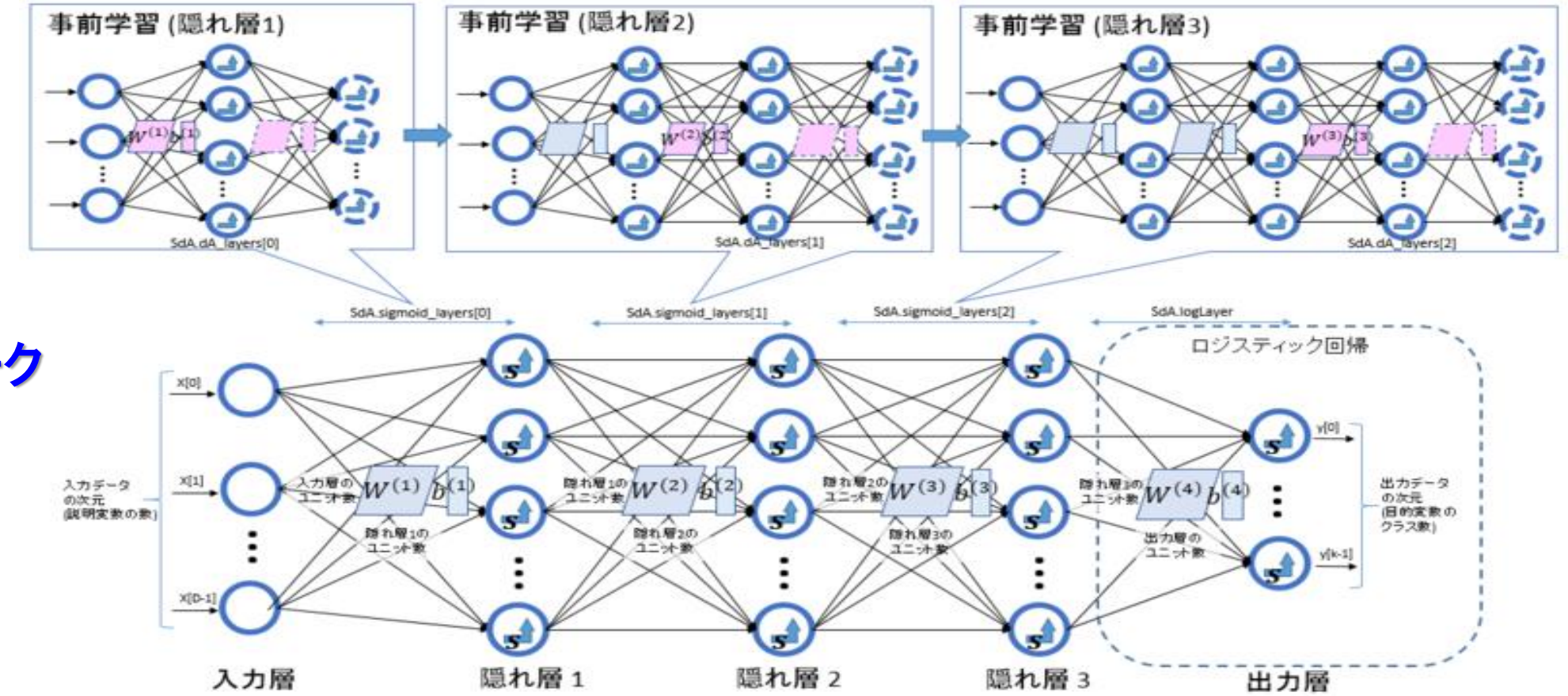


https://nnadl-ja.github.io/nnadl_site_ja/chap6.html

□機械学習型人工知能 Machine learning type artificial intelligence

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）
Execution environment: Development of machine learning methods (deep learning: deep learning)

順伝搬型
ニューラルネットワーク
Forward propagation
neural network



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

□機械学習型人工知能 Machine learning type artificial intelligence

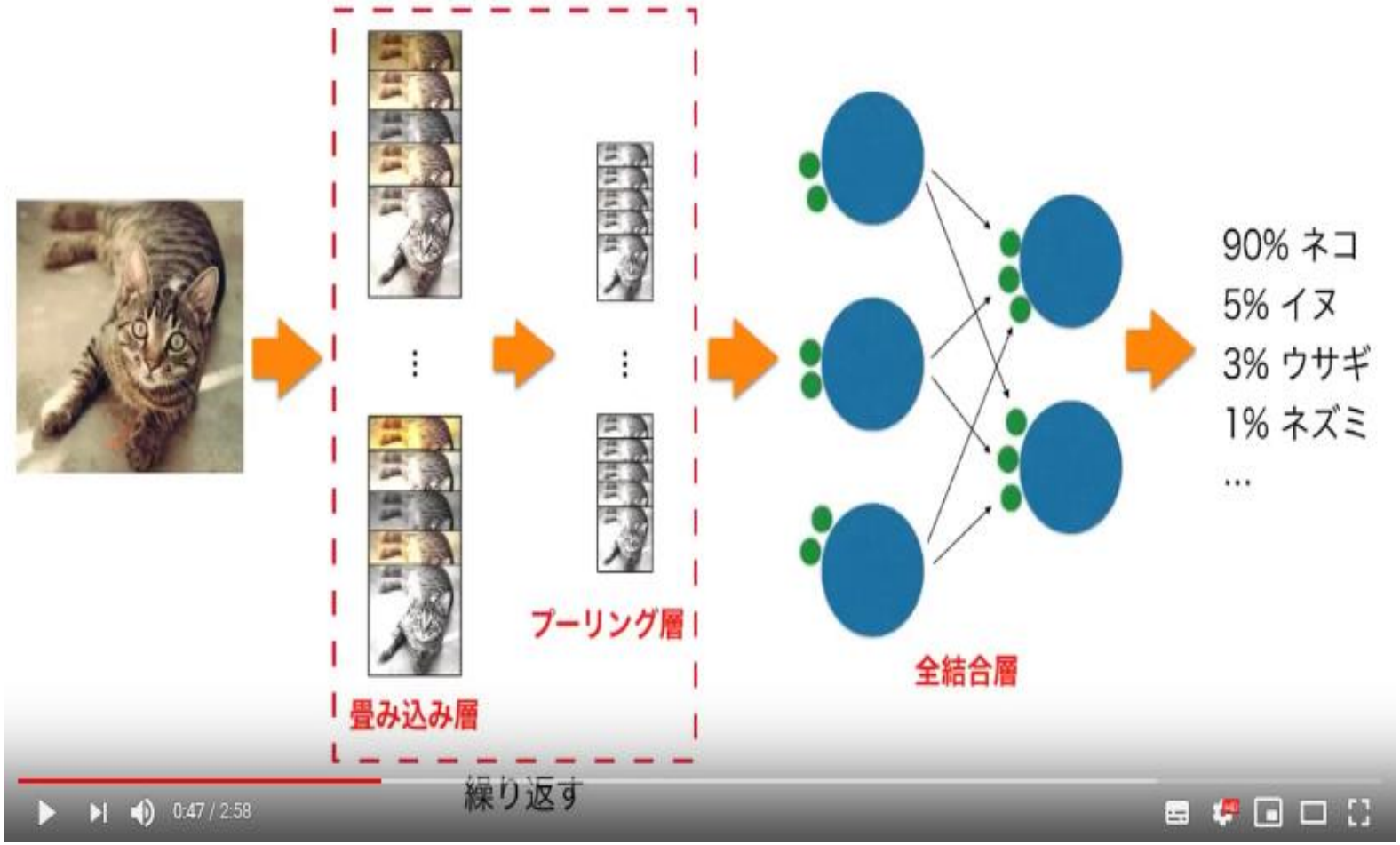
In Silico Data
Miracles by the KY-methods

・開発歴：畳み込みニューラルネットワーク Convolutional neural network

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。

CNN is a neural network that consists not only of all connected layers, but also of convolution layers and pooling layers.

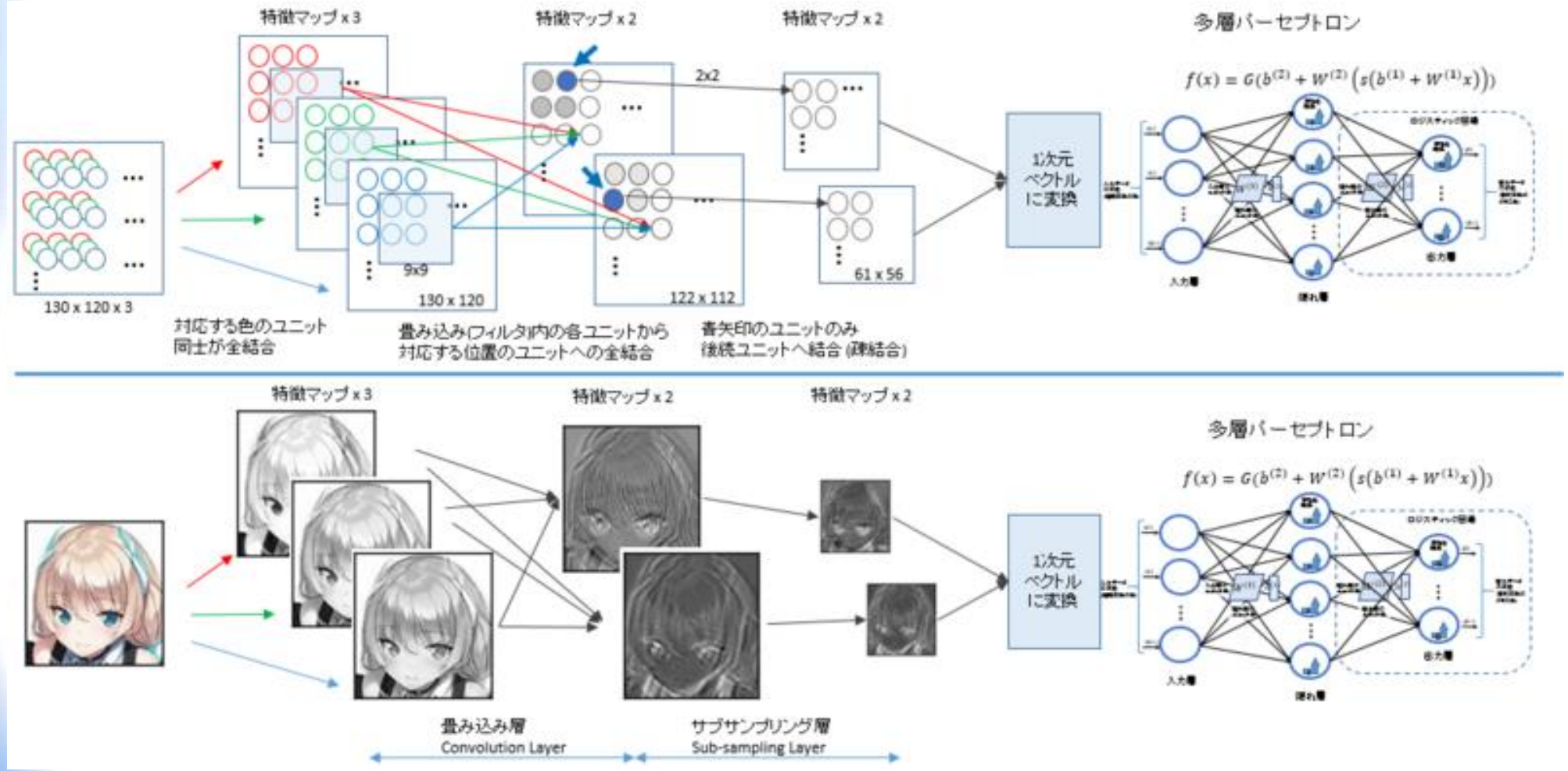


https://www.youtube.com/watch?v=Kr84vlulqTA

口機械学習型人工知能 Machine learning type artificial intelligence

In Silico Data
Miracles by the KY-methods

・開発歴：畳み込みニューラルネットワーク　Convolutional neural network

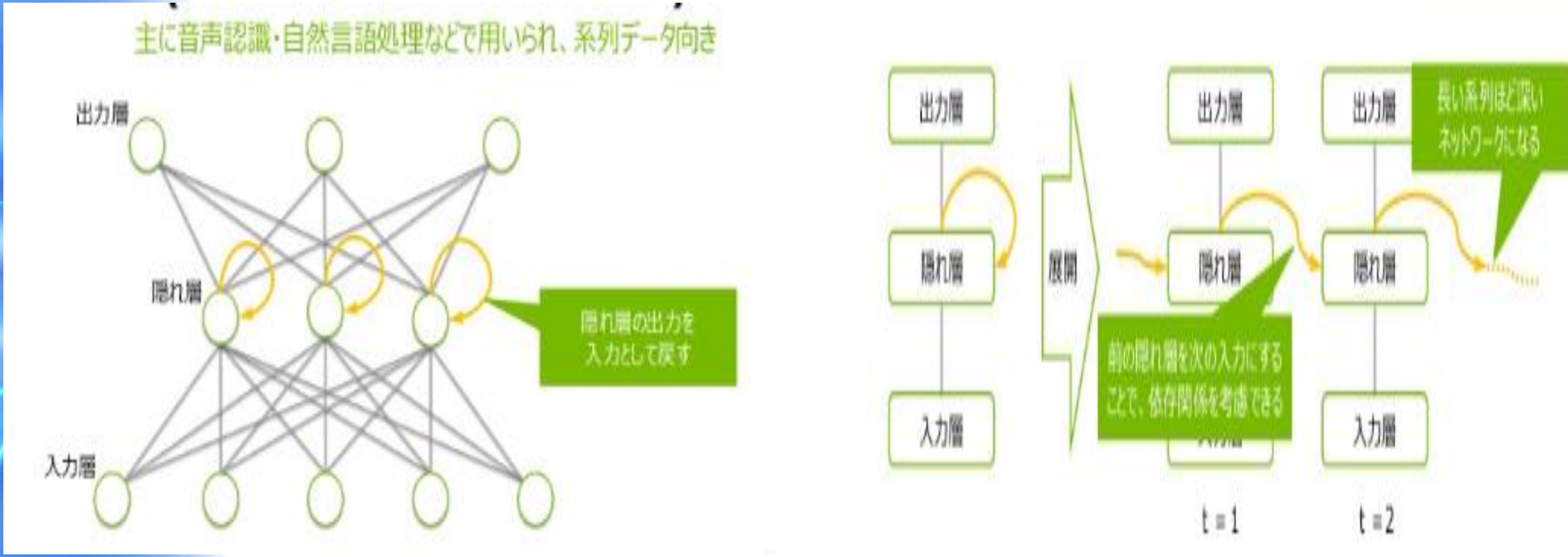


http://sinhrks.hatenablog.com/entry/2014/12/07/203048

□機械学習型人工知能 Machine learning type artificial intelligence

リカレント（再帰型）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
  The output data of the intermediate layer is used again as input data
- 時系列データの解析や言語処理等に利用される
  Used for time-series data analysis and language processing



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

□機械学習型人工知能 Machine learning type artificial intelligence

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

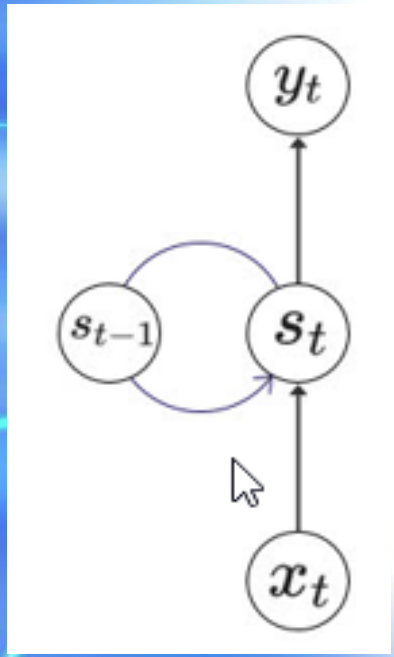
◇リカレントニューラルネットワークには様々なタイプがある

There are various types of recurrent neural networks

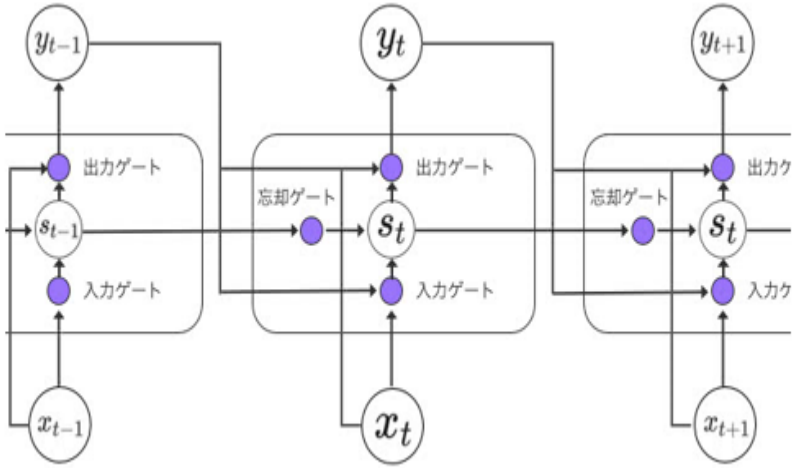
基本の考え方は、自身の出力を改めて入力に使うことである

The basic idea is to use your output again as input

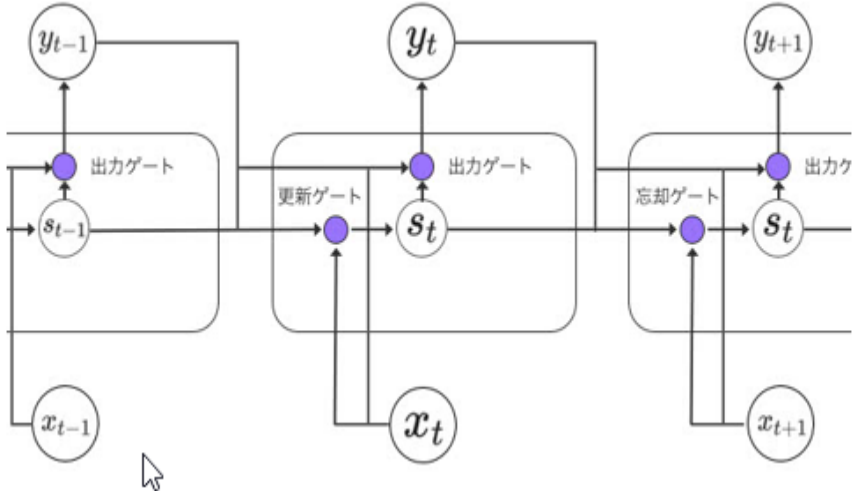
Simple RNN



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 Machine learning type artificial intelligence

In Silico Data
Miracles by the KY-methods

・機械学習型人工知能構築上での留意点（1）
Points to keep in mind when building machine-learning artificial intelligence

◇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）

The number of samples used for learning is extremely large (sample collection in compound-related fields)

・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

・ When deploying as a data analysis method, if the number of samples is small, accidental correlation and overfitting will occur, reducing the reliability of artificial intelligence

◇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

In the learning sample, it is necessary to prevent information from being biased

・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

Sample data bias is dangerous not only for artificial intelligence but also for data analysis

□機械学習型人工知能 Machine learning type artificial intelligence

In Silico Data
Miracles by the KY-methods

・機械学習型人工知能構築上での留意点（2）

Points to keep in mind when building machine-learning artificial intelligence

◇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

Factors cannot be extracted because the network structure is extremely complex

・構造－活性/毒性/物性相関等の研究では、要因抽出が極めて大事である

Factor extraction is extremely important in research on structure-activity / toxicity / physical property relationships, etc.

◇データクレンジング(Data Cleaning)が大事　Data cleansing is important

・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

The sample data for learning should be free from noise in various forms

□機械学習型人工知能 Machine learning type artificial intelligence

In Silico Data
Miracles by the KY-methods

・機械学習型人工知能構築上での留意点（3）
Points to keep in mind when building machine-learning artificial intelligence

◇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）

The number of samples used for learning is extremely large (sample collection in compound-related fields)

・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

When deployed as a data analysis method, if the number of samples is small, accidental correlation and overfitting will occur, reducing the reliability of artificial intelligence

・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。

A neural network has a more complicated network structure than a perceptron or the like.

□機械学習型人工知能 Machine learning type artificial intelligence

In Silico Data
Miracles by the KY-methods

・機械学習型人工知能構築上での留意点(4)
Points to keep in mind when building machine-learning artificial intelligence

・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。

Deep learning has a more complicated network structure than a neural network. For this reason, it is essential to increase the number of samples when using deep learning as a data analysis.

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

Alpha Go, the world's best, has tens of millions of studies in total, including computer-to-computer reinforcement learning

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

In the case of a dramatic increase in image recognition, millions of images are used

・機械学習型人工知能適用上での解決すべき点や対応
Points and solutions to be solved when applying machine learning type artificial intelligence

1. 化合物情報の扱い Handling of compound information
・化合物の画像情報、異性体の扱い、一元一項対応
Compound image information, handling of isomers, one-to-one correspondence

2. サンプル数の問題　Sample number problem
・異性体の扱いによるサンプル数の水増し Increased number of samples by handling isomers
・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
Support for small samples; transfer learning, sparse modeling, etc.

3. 要因説明の問題 Problem of factor explanation
・ネットワークからの情報取り出し　Retrieving information from the network
・ネットワーク構造の単純化　Simplification of network structure

化合物情報の扱いHandling of compound information：　Graph convolution

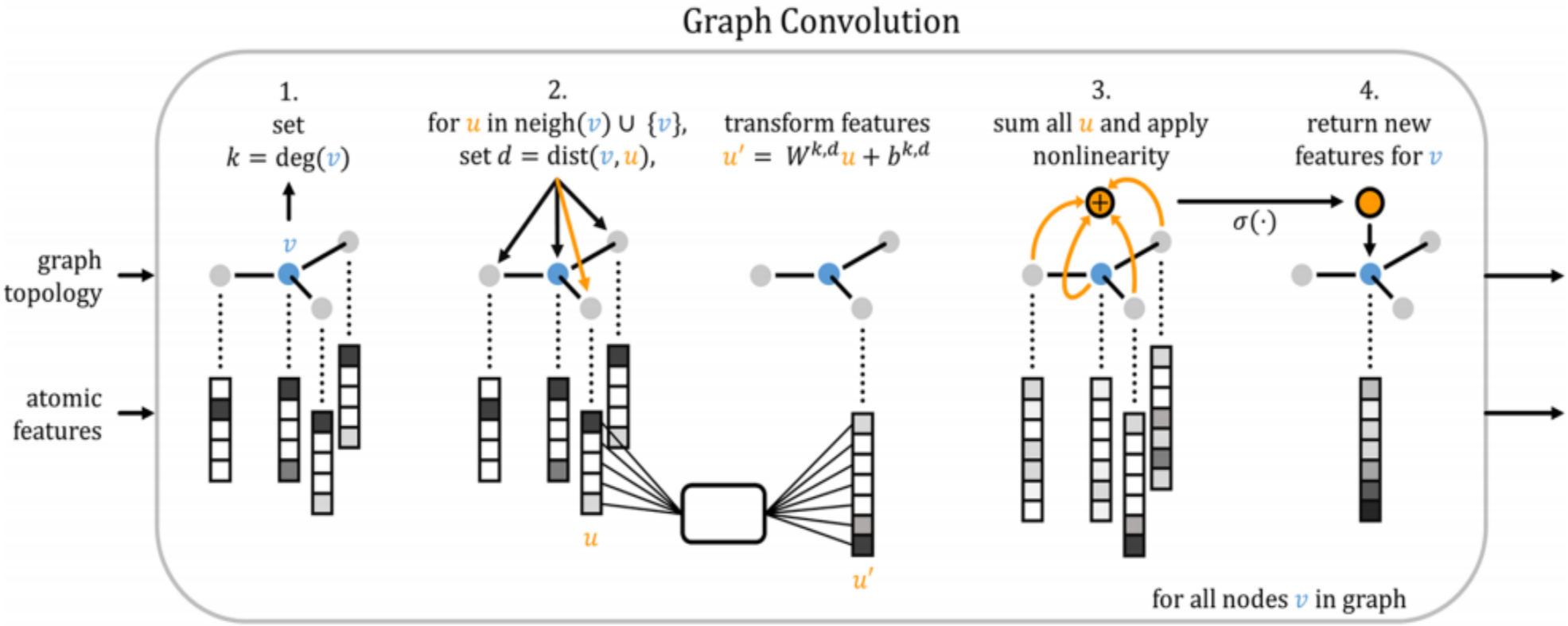
・Graph convolutionは最近のニューラルネットワーク型人工知能の展開過程で、化合物情報を数値データ化する手法としてトポロジー理論を用いて展開された。
Graph convolution was developed by using topology theory as a method to convert compound information into numerical data in the process of recent development of neural network type artificial intelligence.

・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。
Although it is the digitization of compound information by topology, a similar approach has already been developed since the beginning of the structure-activity relationship.

・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。
Housoya index, MCI (Molecular Connectivity Index), and various other indexes have been proposed for quantifying compounds by topology.

・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。
Each index has been energetically developed in research fields such as drug discovery and physical properties, and various pharmacological activities and correlation studies with compound physical properties have been published.

化合物情報の扱いHandling of compound information： Graph convolution



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

化合物情報の扱いHandling of compound information： Graph convolution

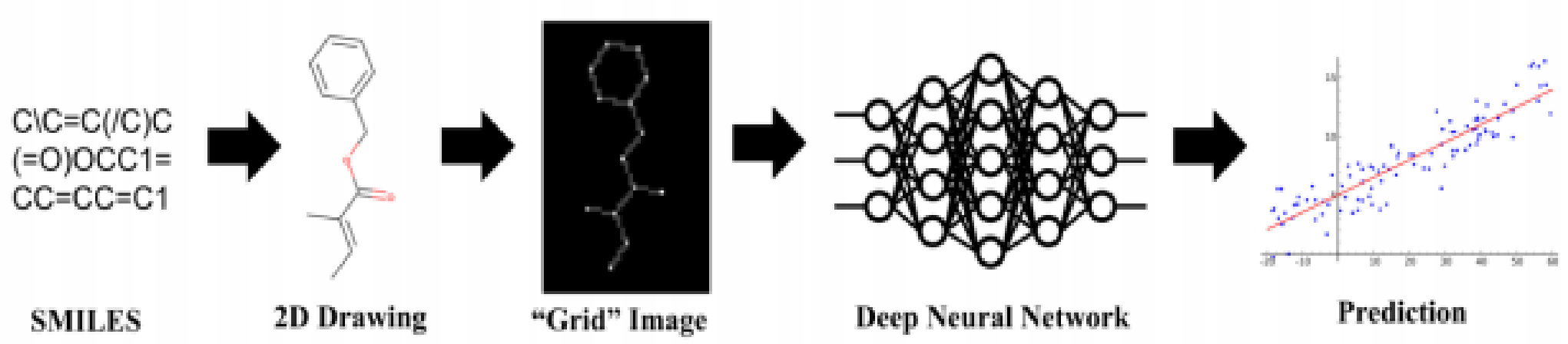
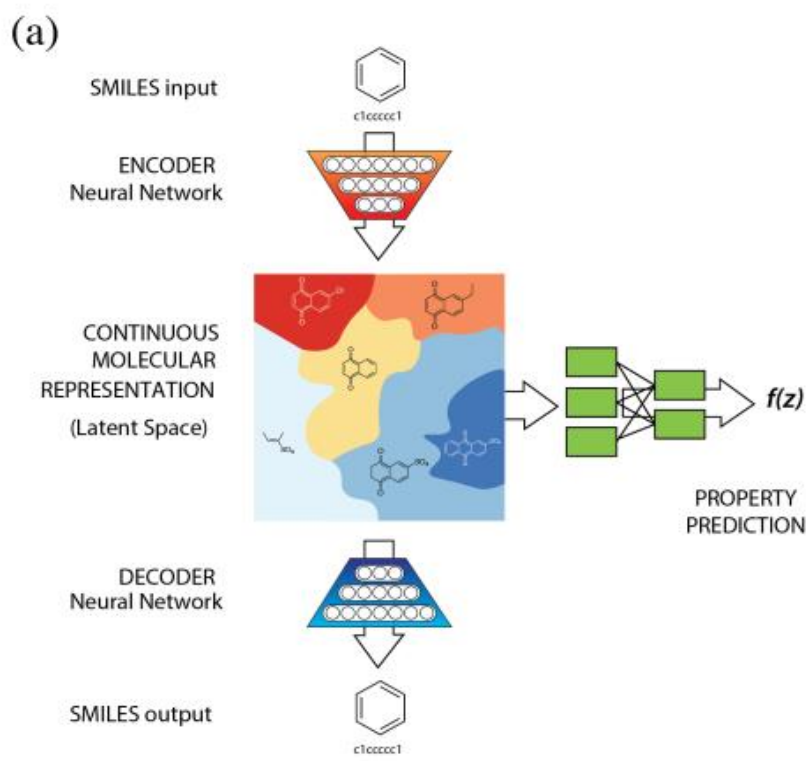


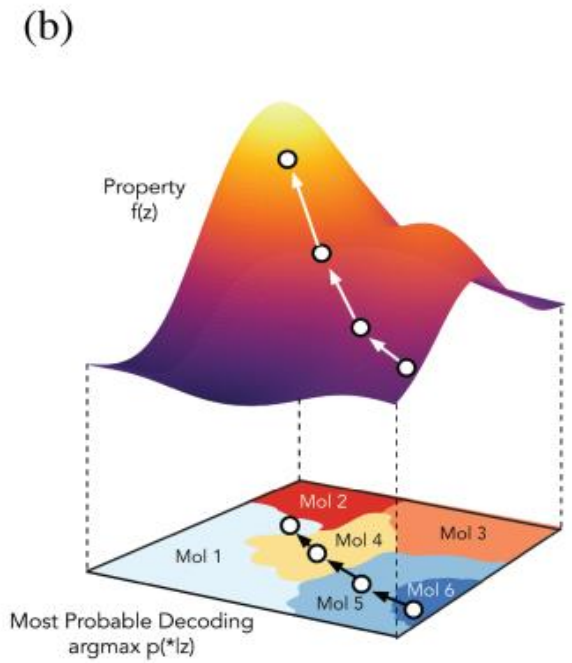
Figure 2: Illustration of the Chemception framework. After a SMILES to structure conversion, the 2D images are mapped onto an 80 x 80 image that serves as the input image data for training a deep neural network to predict toxicity, activity, and solvation properties.

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

化合物情報の扱いHandling of compound information： 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。





Research that brought Deep Leaning to molecular design is [Gómez-Bombarelli +, 2016]. In this research, the SMILES notation, which is the string representation of the molecule, is converted to a real vector using the variational autoencoder (VAE), and the vector optimized by Bayesian optimization is returned to SMILES to design the molecule. The problem with this method is that the rate at which the generated vector is not compatible with the numerator because the optimization vector is returned to SMILES in the VAE space is very high.

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
Notes and summary when applying artificial intelligence in the chemical field

□現在の人工知能による医療や創薬へのアプローチ
Current approaches to medical and drug discovery using artificial intelligence

●WATSONとの連携による医療関連アプローチ
Medical-related approach in cooperation with WATSON
●深層学習採用による創薬関連アプローチ
Drug design related approaches by employing deep learning

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。

At present, many topics are offered, deep learning is trial-like, and it can be said that there are no cases with full-scale success stories.
Although it has been reported that a large number of Journals were entered and a new drug candidate was obtained as a result, there was no report that pharmacological activity actually appeared.

今のAIではできないことも多い

人間には考えられない見間違いも

元の画像（左）に人間には分からない情報（中央）を加えたスクールバス（右）を、AIはダチョウと判断＝グーグルの論文より引用

離れた場所の音声の内容を認識できない

膨大なデータがないと賢くならない

状況に応じた抑揚をつけられない

本や食器は上手につかめない

包丁で野菜を切ることができない

人間の何気ない動きで難しいものも

グラフィックス 佐藤季司

AI、弱点克服へ厚い壁
感情表現や判断力　まだまだ

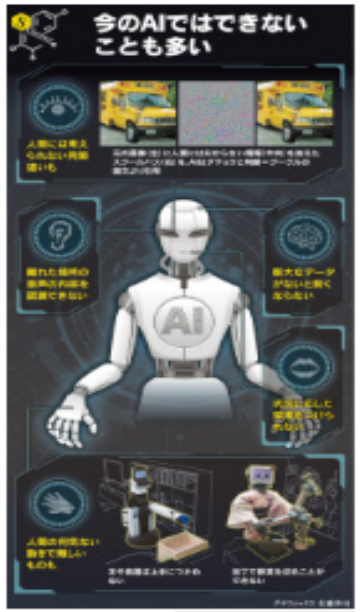
2018. 12. 09(日)

［有料会員限定］

保存 共有 印刷 その他

人工知能（AI）の進歩が著しい。将棋や囲碁で人間をしのぐ性能を示して以降、自動翻訳や医療画像の診断、自動運転など様々な分野で導入機運が高まる。一方で、何が弱点かも明らかになり、次の研究課題になっている。AIは革新を起こす重要な技術だが、決して万能ではない。効用と限界をよく吟味する必要がありそうだ。

人間とうまく対話できる技術の実現は、いまAIに求められている大きな目標だ。しかし研究現場ではいくつかの壁に当たっている。



画像の拡大

「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」

スマートスピーカーやコールセンターの声の聞き取りには、AIが使われている。正しく認識するためにはマイクの近くで話すことが必要だ。広い部屋で行われる会議や記者会見などを、音声認識システム1台で対応し、文字起こしするのは難しい。

その理由は、人間の聞き取りではそれほど支障のない反響音の認識を、AIが苦手としているためだ。特にマイクと話し手の距離が遠いと、部屋中の壁に当たって反響し精度が上がらない。

「僕は状況に合わせて抑揚をつける話し方が下手なんだ。うれしいときに声が上ずったり悲しいときに罵気がなくなったりと、人間のように喜怒哀楽をうまく表現できない」

AIによる話し方は、ニュースを読むアナウンサーのように淡々と読み上げる印象が強い。NTTの宮崎昇主幹研究員は「アクセントは上手につけられるようになったが、表現力はこれから」と話す。人間は対話する相手の感情を予想し状況に合わせながら抑揚を使い分ける。そんな芸当の実現はもう少し先になる。

「僕は賢いと思われているけれど、物覚えがいいだけで、判断力は高くないよ」

現在のAIの中核技術である「深層学習」は、誤った情報や雑音がない膨大なデータを学ばせる手法が主流だ。高い精度の判断には最低でも数千から数十万件のデータが必要といわれる。これに対し、人間は少数のデータで済む。図鑑で犬や猫を学んでおけば、実物を見てすぐに犬か猫かを判断できる。人工知能学会の浦本直彦会長は「少ないデータで判断するのは人間の方が得意」と話す。

深層学習を用いたAIが注目を集めたきっかけは画像認識の精度が格段に向上したためだ。「機械が目を手に入れた」と例えられている。そこに意外な落とし穴があった。

「人間には考えられない見間違いを起こすこともあるよ」

グーグルの研究者らは人間にはスクールバスに見える画像を、AIがダチョウと判定した事例を論文で公表している。元のスクールバスの画像に人間には分からない加工を施すと、AIの判定が変わってしまった。

自動運転の研究にとってこれは都合が悪い。この仕組みを悪用されると、例えば「止まれ」の道路標識が「60キロ制限」に誤って認識されてしまう恐れもある。人間とAIでは画像からつかむ特徴が異なるために起きる問題で、根本的な解決策はまだ見つかっていないという。

「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」

AIベンチャー企業のプリファード・ネットワークス（東京・千代田）は10月、散らかった部屋を片付けるロボットを公開した。ペンや靴下など数百種類の物を認識し、決められた場所に戻す性能を備える。ただ、すべての物を拾い上げられるわけではない。エンジニアの羽鳥潤氏は「本や食器はまだ上手につかめない」と話す。

ロボットの指は人間ほど繊細に動かせず、ある程度の重さがあり平らな本や皿の下にうまく滑り込ませられない。人間が上手に持てる物でもうまく扱えず、AIは不得手である事実を認識できない。

中京大学の橋本学教授らが開発したお茶をたてるロボットは、スプーンで抹茶の粉を取りひしゃくでお湯をすくうなど様々な動作が可能だ。しかし包丁を使って野菜を切ることはできない。「切るときに大きな力が要るし、刃の動きを厳密に制御する必要がある」（橋本教授）

ロボットの頭脳にAIはなくてはならない。融合する研究はこれから大きな流れになると期待されている。ただ難問も多く、AIだけロボットだけの技術開発からは、よい解決策が見いだせない可能性もある。研究者は融合研究を活発にして厚い壁を越えようと考えている。

（大越優樹）

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

In Silico Data
Miracles by the KY-methods

• 機械学習型人工知能の有する基本的弱点
Fundamental weaknesses of machine learning type artificial intelligence

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

本文中：感情表現や判断力まだまだ　Emotional expression and judgment

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」
“My“ ears ”can't recognize distant speech. Speak near the microphone.”

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」
”I'm not good at speaking with inflection in accordance with the situation .... I can't express emotions like humans well.“

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」
“I think I'm smart, but I just remember things and not judge.”

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」
“It can make mistakes that humans cannot think of.”

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」
“Human can use his hands to do various tasks dexterously, but I have a hard time even grabbing a book.”

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

In Silico Data
Miracles by the KY-methods

• 機械学習型人工知能の有する基本的弱点
Fundamental weaknesses of machine learning type artificial intelligence

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

図中：今のAIではできないことも多い　There are many things that AI can't do

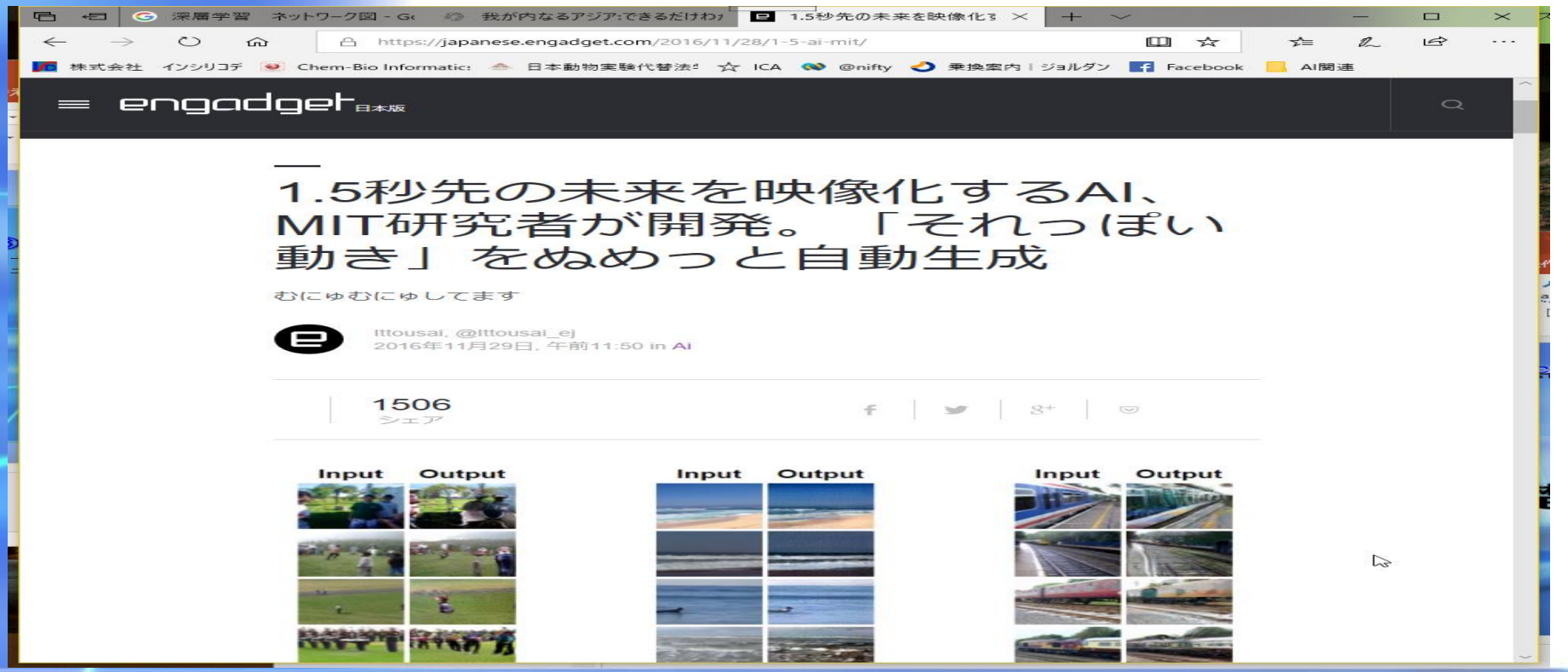
1. 人間には考えられない見間違いも　Misunderstandings that humans cannot think of
2. 離れた場所の音声の内容を認識できない　Unable to recognize remote audio content
3. 膨大なデータがないと賢くならない　You won't be smart without a lot of data
4. 状況に応じた抑揚をつけられない　Can't give inflection according to the situation
5. 人間の何気ない動きで難しいものも　Things that are difficult due to human casual movement

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

In Silico Data
Miracles by the KY-methods

・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
＊**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



・Difficult to apply to things other than learning and acquired information
⇒ Cannot recognize state after 1 second ⇒ Cannot predict after 1 second from photo analysis of moving car ⇒ Learning to recognize moving and non-moving objects is required
＊Even if the number of samples is sufficient, there are things that cannot be acquired by learning

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
Notes and summary when applying artificial intelligence in the chemical field

・機械学習型人工知能の有する基本的弱点
Fundamental weaknesses of machine learning type artificial intelligence

◇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）

The number of samples used for learning is extremely large
（sample collection in compound-related fields）

・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

When deployed as a data analysis method, if the number of samples is small, accidental correlation and overfitting will occur, reducing the reliability of artificial intelligence

◇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

In the learning sample, it is necessary to ensure that the information is not biased

・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

Sample data bias is dangerous not only for artificial intelligence but also for data analysis

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

In Silico Data
Miracles by the KY-methods

・機械学習型人工知能の有する基本的弱点
Fundamental weaknesses of machine learning type artificial intelligence

◇ネットワーク構造が極めて複雑なので**要因抽出**ができない

Factors cannot be extracted because the network structure is extremely complex

・構造－活性/毒性/物性相関等の研究では、要因抽出が極めて大事である

Factor extraction is extremely important in research on structure-activity / toxicity / physical property relationships, etc.

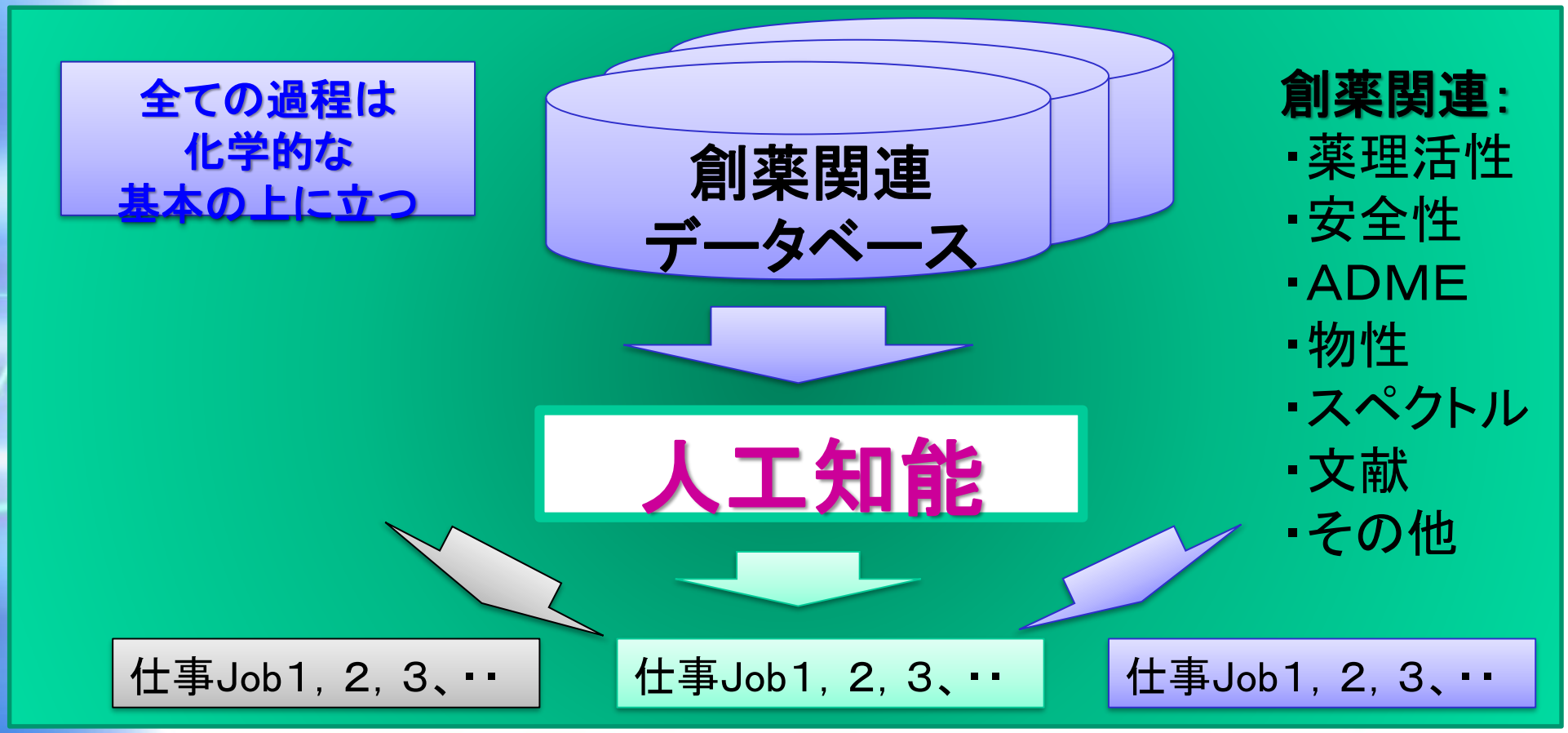
◇**データクレンジング**(Data Cleaning)が大事　Data cleansing is important

・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

The sample data for learning should be free from noise in various forms

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

- 人工知能適用における化学的問題
  Chemical problems in artificial intelligence applications.



□機械学習型人工知能 Machine learning type artificial intelligence

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

Drug discovery is a world centered on compound structural formulas

The chemical researcher's thinking process is considered by the compound structural formula, communicates with each other, and answers with the compound structural formula.

It is important that the artificial intelligence system can interact with the researcher who is the user in the compound structural formula

Example: Drug researcher
What part of the compound structural formula should be changed and how to strengthen the pharmacological activity? ⇒Needs dialogue with researchers

Just like chess, shogi, and samurai, you can't specify the coordinates on the board. Also, it's not just to win

□人工知能適用における化学的問題
Chemical problems in artificial intelligence applications.

□化合物構造式に始まり、化合物構造式に終わる

Begins with compound structural formula and ends with compound structural formula

・研究者の思考過程は総て化合物構造式で終始する

Researchers' thought processes all start with compound structural formulas

・**化合物の表現の問題：　Compound expression problems**

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。

The same compound takes various forms depending on the expression system, such as compound name, molecular formula, two-dimensional structural formula, three-dimensional structural formula, etc., and the contents and amount of information of each expression system are different.

・**入力の問題：Input issues**

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。

Even if there are a huge amount of journals and general chemical literature, it is necessary to accurately input compound structure information necessary for learning artificial intelligence.

・**結果の問題：Result issues**

結果が出たら、人工知能情報の化学情報への変換が重要

Once the results come out, it is important to convert artificial intelligence information to chemical information

□人工知能適用における化学的問題
Chemical problems in artificial intelligence applications.

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

Example: “One-way multiplet” problem for compounds
・Handling compounds that have entered artificial intelligence with multiple faces?
Require function to check that they are the same compound
・Possibility of being judged as a different compound in the learning process

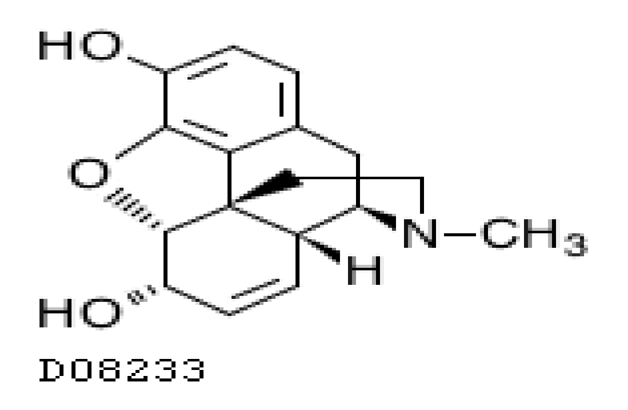
例：Journal情報利用上での問題

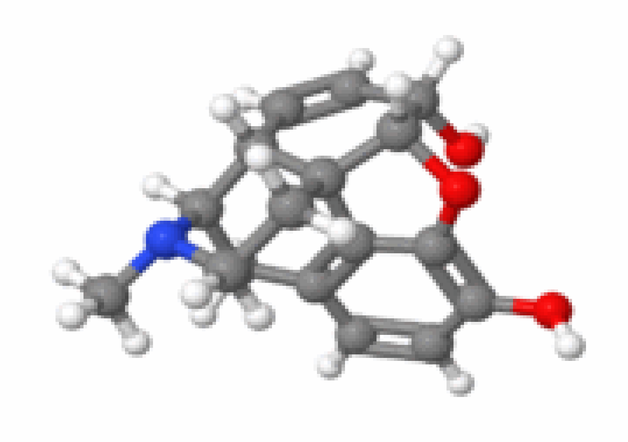
- 化学やバイオ関連分野の論文は基本的に成功事例
  成功のみ掲載されている。このような成功事例のみを
  学習した結果提案される化合物は、
  成功／失敗化合物？　→　失敗というフィルターがない
- 入力Journal数は精度の保証にならない
  数が多いほど上記の偏向学習が進んでいることの証拠

Example: Problems using Journal information
・Chemical and bio-related fields are basically successful cases
Is the compound proposed as a result of learning only successful examples a success / failure compound? → No filter for failure
・Input Journal number does not guarantee accuracy
Evidence that the above-mentioned biased learning is progressing as the number increases

□人工知能適用における化学的問題
Chemical problems in artificial intelligence applications.

化学者の思考過程で必須
Essential in the chemist's thinking process





化学者はイメージも議論も出来ない

Molファイル:

```
 22 26  0  0  1  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  0  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  0  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
 15 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

□人工知能適用における化学的問題
Chemical problems in artificial intelligence applications.

＊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

- Information that chemists can imagine is compound structural formulas, which cannot be discussed only with numbers and letters.

＊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

- Computers can handle numbers and character codes, not structural image information.

**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**

Necessary to fill the gap between the above two events when implementing artificial intelligence

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

In Silico Data
Miracles by the KY-methods

・機械学習における問題 Problems in machine learning

□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

□ Recent artificial intelligence is mainly machine learning

advantage:

・Can handle large amounts of data
・Contents that could not be expanded with artificial intelligence can be expanded.
・No need for know-how (rules)
Develop in fields where there is no know-how
・It may be possible to discover new knowledge

Disadvantages: problems

・Can the system understand chemical knowledge?
・Result feedback is technically difficult
・It is difficult to make new knowledge into a level that humans can interpret

・機械学習における問題 Problems in machine learning

□深層学習実施上での留意点 Points to keep in mind when conducting deep learning

1. 学習に用いるサンプル数問題→過剰適合の回避
   Number of samples used for learning → avoiding overfitting
   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
   If the network hierarchy is deep, the number of samples required for learning increases rapidly.
2. 学習の偏り回避→サンプルの学習内容が大事
   Avoid learning bias → Learning content of the sample is important
3. 現在は画像／音声／文字認識が主体のネットワーク構成
   Currently, the network configuration mainly consists of image / sound / character recognition
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
   Even if the result is good, it is extremely difficult to extract factor information from the network.

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

In Silico Data
Miracles by the KY-methods

・生命科学分野における今後の人工知能の展開（1）
Development of future artificial intelligence in the life science field

□展開分野　Deployment field

歴史的に化学生物分野での人工知能展開事例は多い

Historically, there are many cases of artificial intelligence deployment in the field of chemical biology

□実施手法　Implementation method

現実の条件に即した機械学習の改良／開発

Improvement / development of machine learning according to actual conditions

□ハイブリッド型　Hybrid type

・機械学習およびルールベースMachine learning and rule base

・多変量解析／パターン認識との連携Cooperation with multivariate analysis / pattern recognition

□Iotや分析／医療機器との連携　Collaboration with IoT and analysis / medical devices

□その他の展開　Other developments

◆化学分野で人工知能を適用する時の注意とまとめ
Notes and summary when applying artificial intelligence in the chemical field

・生命科学分野における今後の人工知能の展開（1）
Development of future artificial intelligence in the life science field

□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

□ Is it possible to use large amounts of data efficiently?

・ If the data format and information contents are not complete, it is difficult to handle with machine learning, and it is important to organize the data for artificial intelligence

・ Even if there is a quantity, information necessary for the purpose of artificial intelligence cannot be extracted, or biased information cannot be used for learning

□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

□ Machine learning technology for accurate understanding of chemical structural formulas

・ It is expected that it will be difficult to automatically learn various problems specific to chemistry from a large amount of data.

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

When targeting compounds such as drug discovery and safety, a comprehensive approach not only with machine learning but also with existing know-how and multivariate analysis / pattern recognition (chemometrics) technology The most reasonable to think about

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