

# 機械学習と機械発見

## データ中心型の化学・材料科学の教訓とこれから

---

2021年10月26日

瀧川 一学

ichigaku.takigawa@riken.jp

理化学研究所 革新知能統合研究センター  
iPS細胞連携医学的リスク回避チーム



# 自己紹介：瀧川 一学 (たきがわ いちがく)

## 専門：機械学習と機械発見

特に離散構造を伴う機械学習 + データ中心的な自然科学研究

現在の主業務：**幹細胞生物学(理研) + 化学(北大)**

**10年 北大** 工学研究科 システム情報工学専攻 博士課程修了  
**(1995~2004)** "劣決定信号源分離のL1ノルム最小解の理論分析"

**7年 京大** 化学研究所 バイオインフォマティクスセンター 助教  
**(2005~2011)** 薬学研究科 医薬創成情報科学専攻 (兼務)

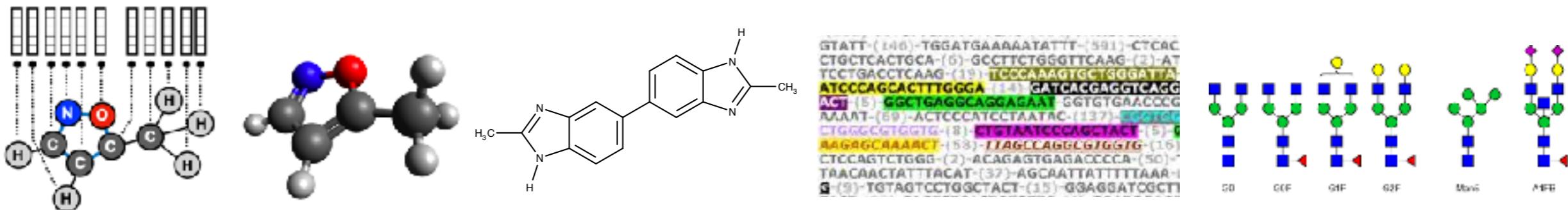
**7年 北大** 情報科学研究科 情報理工学専攻 准教授  
**(2012~2018)** JSTさきがけ 材料インフォマティクス領域 (兼務)

**?年 理研(京都)** AIPセンター iPS細胞連携医学的リスク回避チーム 研究員  
**(2019~)** 北大 化学反応創成研究拠点 (クロスマポイント)

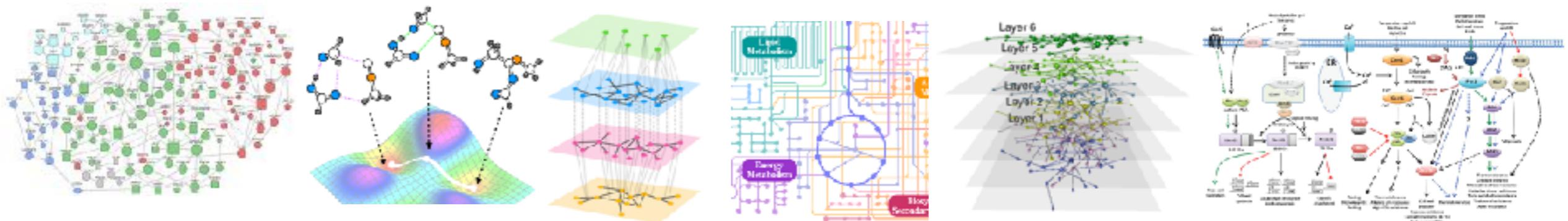
# 離散構造を伴う機械学習

集合、論理、関係、組合せ、系列、木、グラフ、代数系、言語、…

- 対象が「離散構造」を持つ

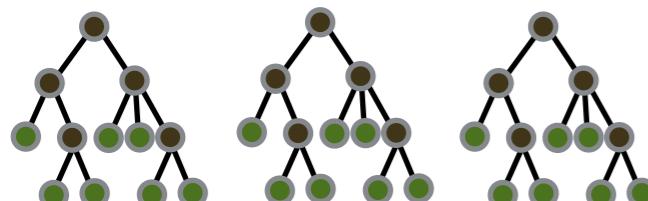


- 対象の関係が「離散構造」を持つ

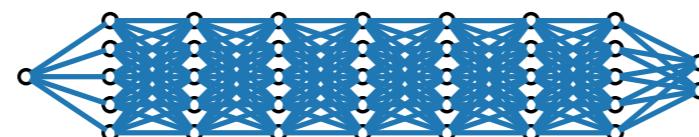


- モデルが「離散構造」を持つ

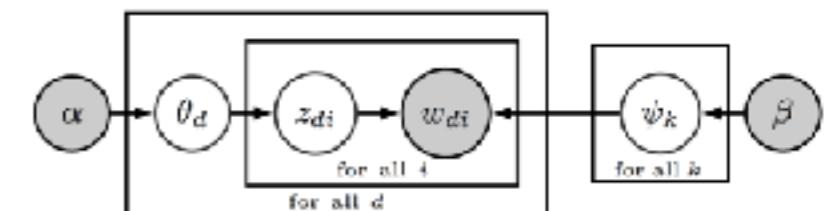
決定木・決定DAG



ニューラルネットワーク



確率的プログラミング



# 今日のテーマ

- **自己紹介 (機械学習と自然科学の境界)**
- **機械学習とは新しいプログラミングの方法**
- **機械学習屋は一体何が楽しいのか？**
  - 分子の表現と機械学習
  - グレイボックス最適化：演繹 + 帰納(論理推論と統計的予測)の統合
- **自然科学研究で機械学習を使おうとすると必ずぶつかる本当に難しい問題**
  - The Two Cultures：データモデリングと予測アルゴリズム
  - 予測か理解か：Rashomon効果, Underspecification, 解釈多様性
  - 人間の認知バイアスに由来する問題：仮説、失敗、成功バイアス、etc.
- **機械学習から機械発見へ**
  - 「発見」「理解」の道筋は合理化できるのか？自動化できるのか？

転職：2019年4月1日～

北海道大学情報科学研究科の研究室をcloseし下記2組織の  
「クロスマポイントメント」へ

- 理化学研究所  
革新知能統合研究センター (AIP)  
iPS細胞連携医学的リスク回避チーム 研究員
- 北海道大学  
化学反応創成研究拠点 (WPI-IICReDD) 特任准教授



# 文科省 世界トップレベル拠点形成プログラム(WPI)

[https://www.mext.go.jp/a\\_menu/kagaku/toplevel/](https://www.mext.go.jp/a_menu/kagaku/toplevel/)

## コンセプト

WPIプログラムは、「世界最高レベルの研究水準」、「融合領域の創出」、「国際的な研究環境の実現」、「研究組織の改革」という4つのミッションの下、ブレインサーチュレーションの中にしっかりと位置づけられる、「目に見える研究拠点」の形成を目指しています。WPIセンターは、日本の研究機関のモデルになるとともに、科学技術に革新をもたらすことを期待されています。



## 特徴



「目に見える研究拠点」の実現へ

# 世界トップレベル拠点形成プログラム(WPI)と情報科学

## 世界トップレベル研究拠点

宇宙/地球・生命/知性  
の起源



平成19年度採択



平成24年度採択



平成29年度採択

生物



平成19年度採択



平成19年度採択



平成24年度採択



平成24年度採択



平成30年度採択



平成29年度採択

材料/エネルギー



平成19年度採択



平成19年度採択



平成22年度採択



平成30年度採択

データ・情報科学



International Research Center for Neurointelligence

ニューロインテリジェンス国際研究機構

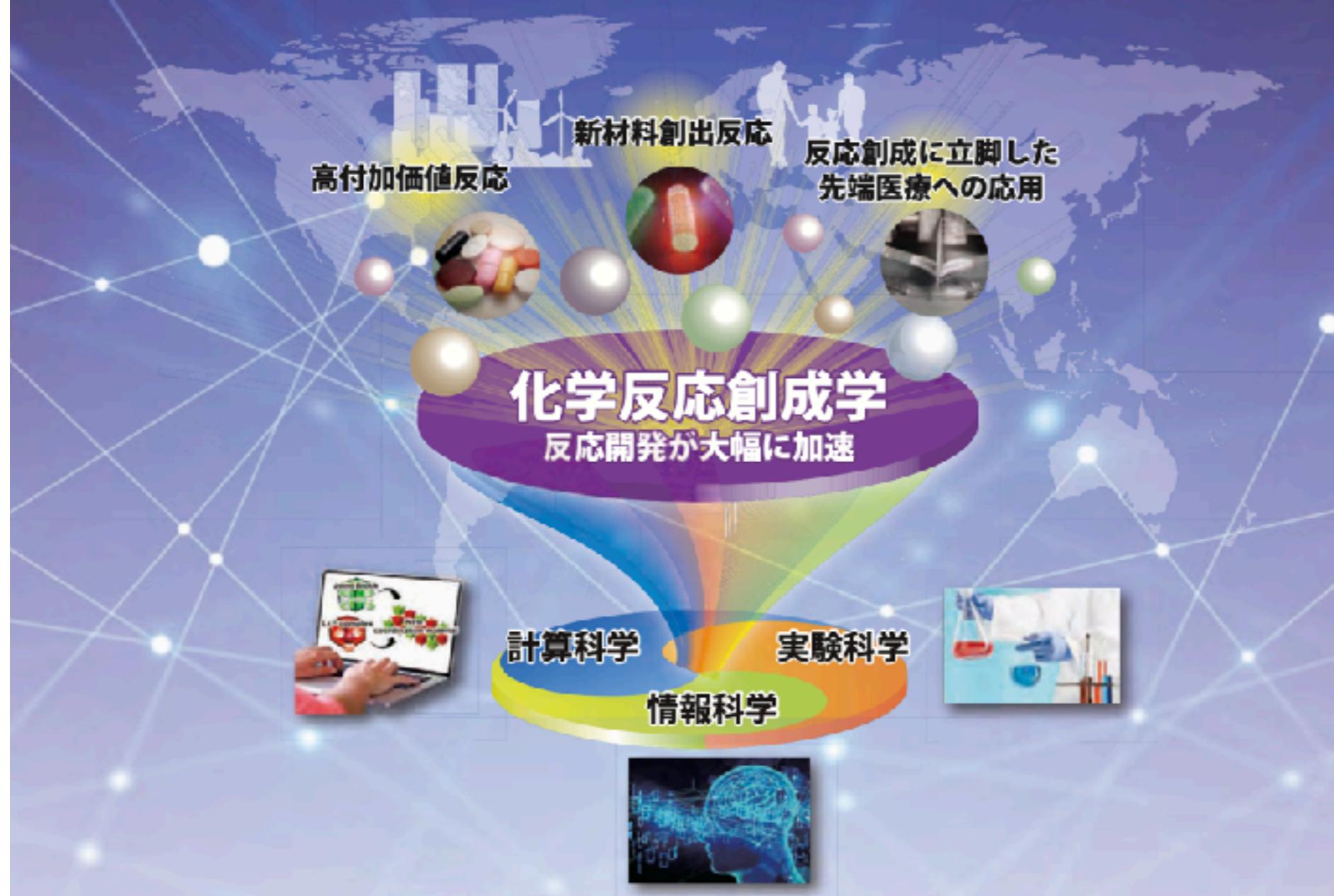


データ・情報科学に重なる  
WPI拠点は二拠点のみ

# 北海道大学 化学反応創成研究拠点 (WPI-ICReDD)

## 化学反応創成研究拠点 (ICReDD) とは

化学反応創成研究拠点 (ICReDD / アイクレッド) では、計算科学、情報科学、実験科学の 3 分野を融合させることにより、新しい化学反応をより深く理解し効率的に開発することを目指しています。



# 北海道大学 化学反応創成研究拠点 (WPI-ICReDD)

<https://www.icredd.hokudai.ac.jp>

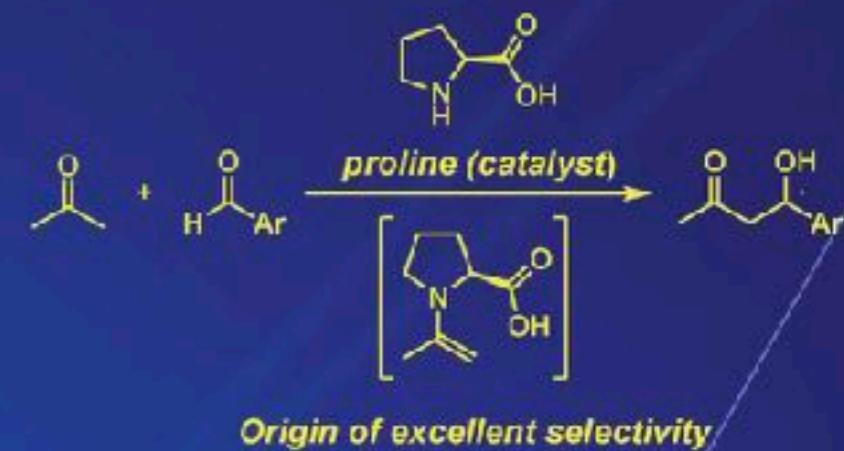


# ノーベル化学賞 2021

## The Nobel Prize in Chemistry 2021 for the Development of Asymmetric Organocatalysis



Nobel Prize Winner  
**Professor Benjamin List**  
ICReDD Principal Investigator



<https://www.youtube.com/watch?v=clvA49BobsI>

Dr. Benjamin List's Thoughts and Message after Winning the 2021 Nobel Prize in Chemistry

Video teleconference with

Benjamin

List

The 2021 Nobel Prize  
in Chemistry laureate



# 理化学研究所 革新知能統合研究センター

<https://aip.riken.jp/>

## センター紹介

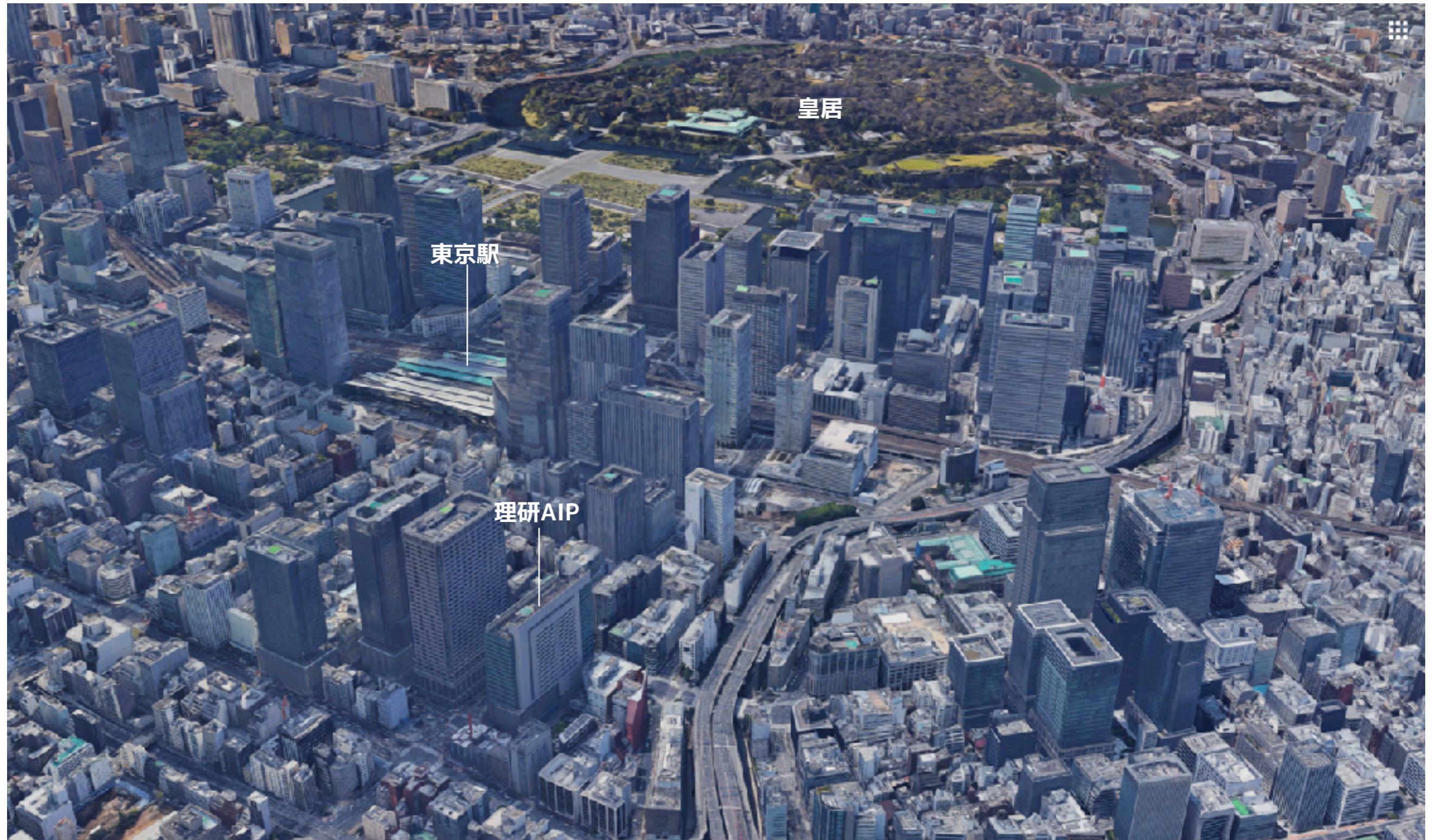
理化学研究所 革新知能統合研究（AIP）センターは、文部科学省AIPプロジェクトの研究拠点として2016年度に設置されました。新たに日本橋オフィスを開設して2017年度より本格的に活動を開始。日本のAI研究をリードすべく活動を行っています。

- 汎用基盤技術研究グループ
- 目的指向基盤技術研究グループ
- 社会における人工知能研究グループ

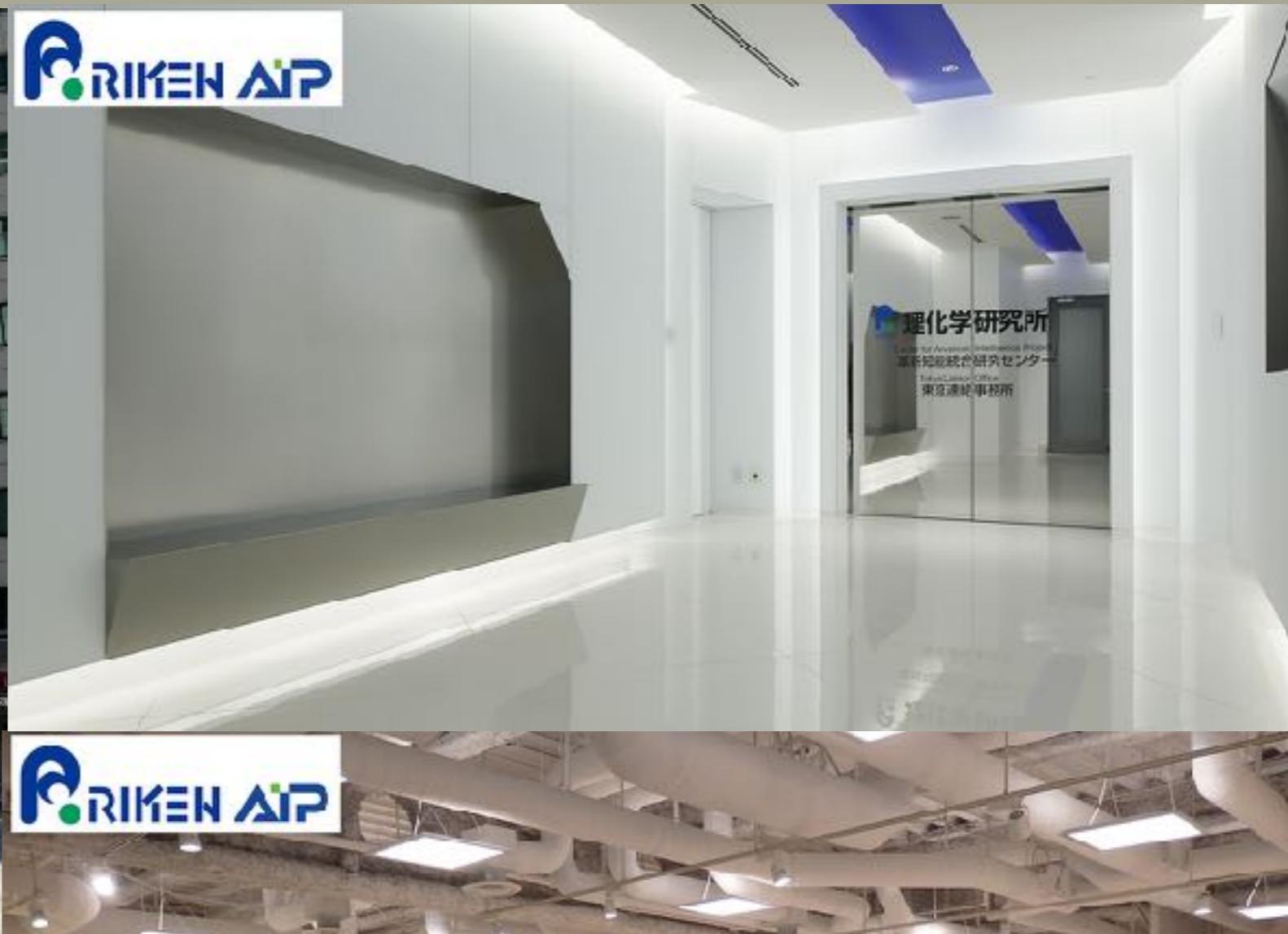
の3つの研究グループを設置し、様々な企業・大学・研究所・プロジェクトと連携して、次の5つの事業を推進しています。

- 基盤技術の開発  
深層学習の仕組みの解明、新しい原理に基づく次世代知能技術の創生
- サイエンス研究の加速  
再生医療、素材開発、ものづくりなど、日本が高い国際競争力を持つ分野を人工知能技術により更に強化
- 社会問題の解決  
高齢者ヘルスケア、防災・減災、インフラ管理などの重要課題に取り組むプロジェクトを人工知能技術で支援
- 人工知能の倫理的・法的・社会的課題の分析  
人工知能技術を日常生活に浸透させていく上で必要となる倫理規準や法制度を議論
- 人工知能研究者・データサイエンティストの育成  
産業界の技術者や学術界の学生・研究員の技術レベルの向上に貢献し、諸外国の大学・研究所との連携体制を構築

# 理化学研究所 革新知能統合研究センター



# 理化学研究所 革新知能統合研究センター



# 勤務地：京阪奈地区(京都府相楽郡精華町)

## けいはんな地区

<https://www.kobe.riken.jp/about/map/keihanna/>



株式会社国際電気通信基礎技術研究所

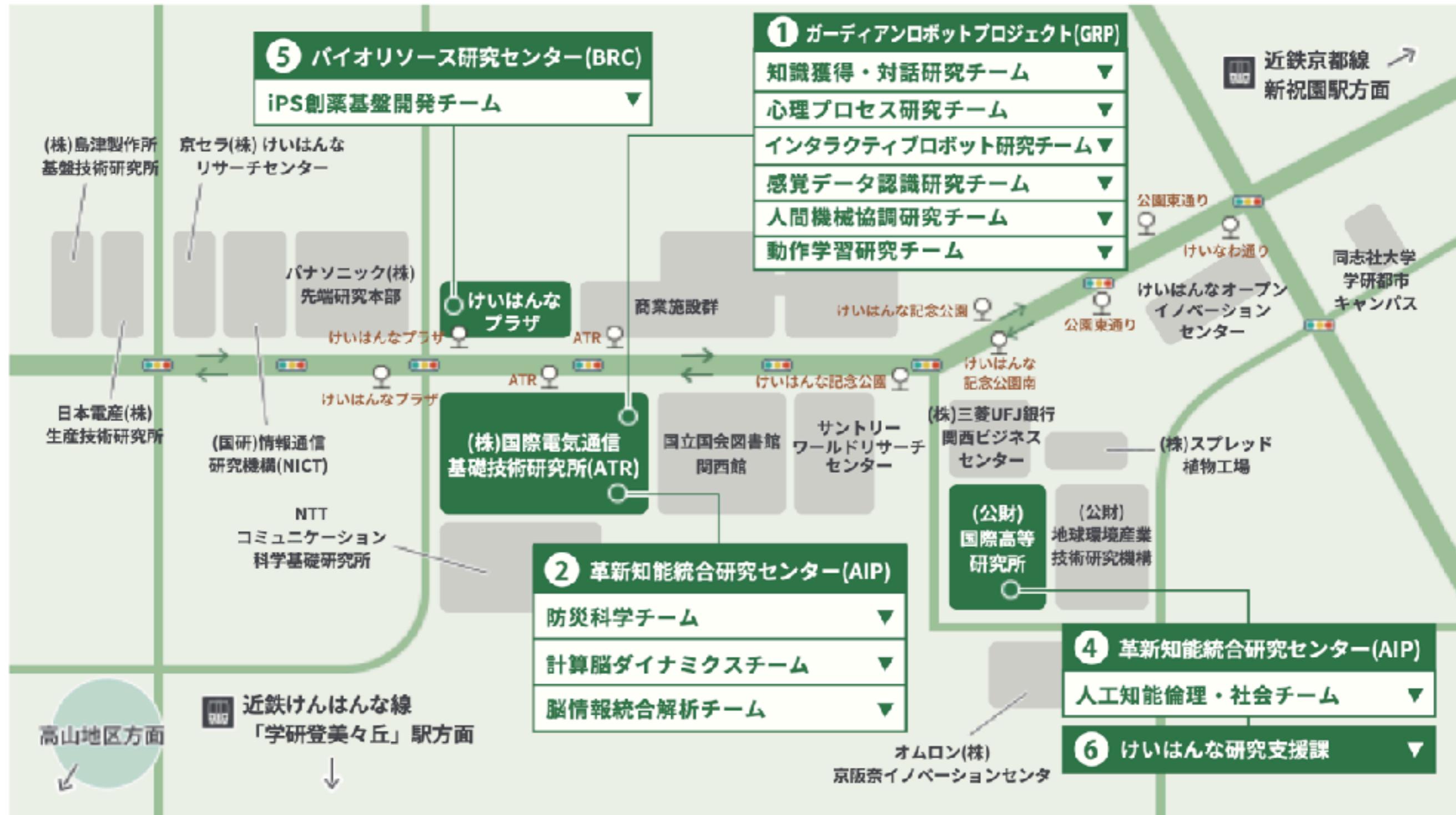


国立大学法人奈良先端科学技術大学院大学



公益財団法人国際高等研究所

# 勤務地：京阪奈地区(京都府相楽郡精華町)



# 勤務地：京阪奈地区(京都府相楽郡精華町)



# 国際電気通信基礎技術研究所 (ATR)

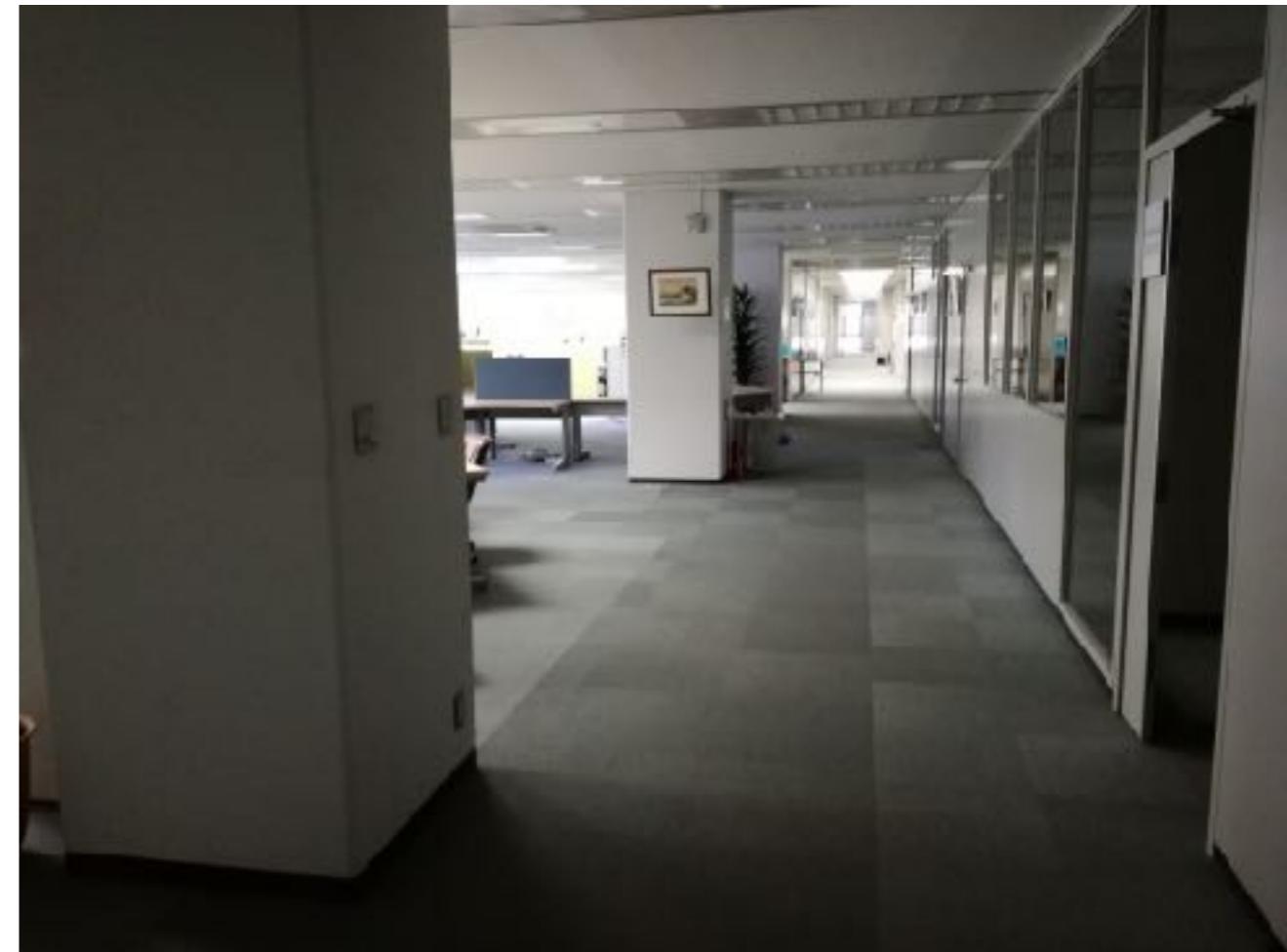
<https://www.atr.jp>



- 理化学研究所  
革新知能統合研究センター (AIP)  
ガーディアンロボットプロジェクト (GRP)
- ATR脳情報通信総合研究所  
脳情報研究所  
認知機構研究所  
脳情報解析研究所

- 深層インタラクション  
インタラクション技術バンク, インタラクション科学研究所, 石黒浩特別研究所, 萩田紀博特別研究所
- 無線・通信  
適応コミュニケーション研究所, 波動工学研究所
- 生命科学  
佐藤匠徳特別研究所

# 国際電気通信基礎技術研究所 (ATR)



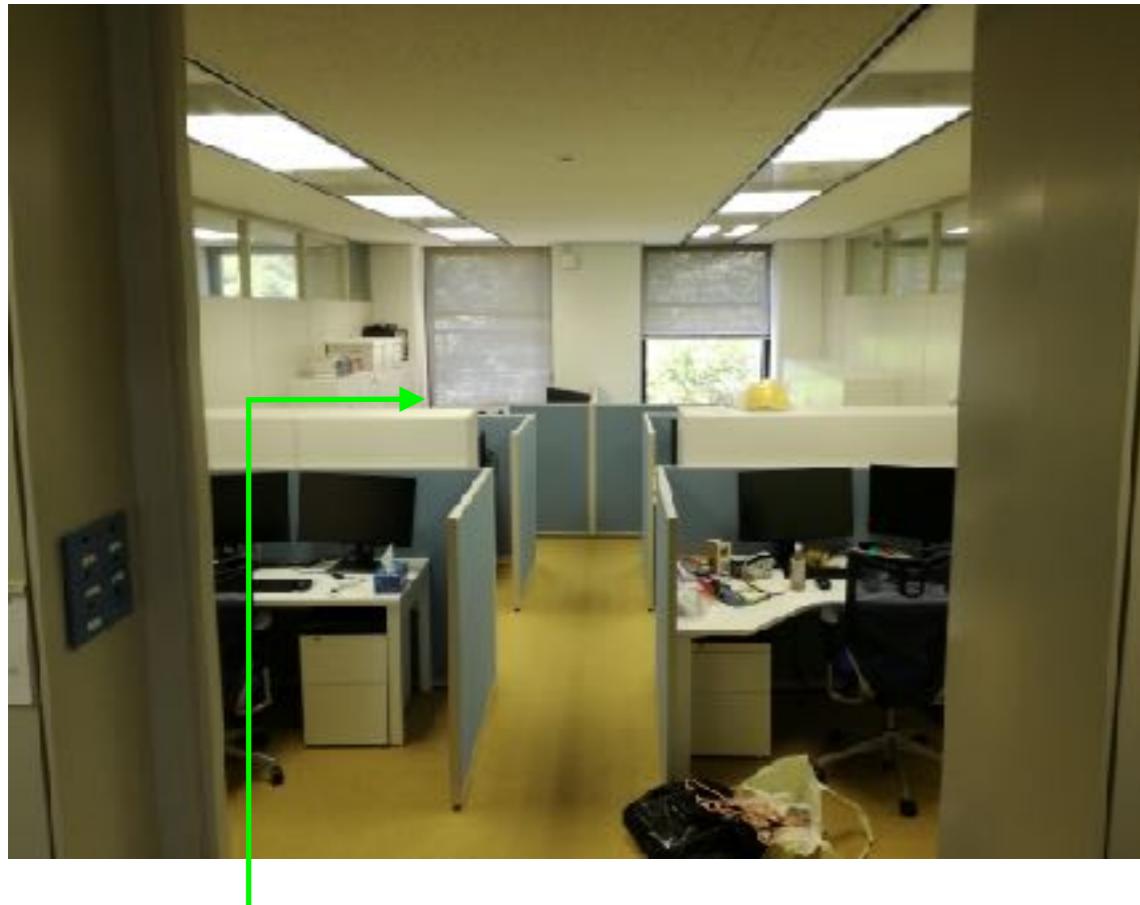
## • 脳情報通信総合研究所

- 脳情報研究所 (CNS)
- 認知機構研究所 (CMC)
- 脳情報解析研究所 (NIA)



- 動的脳イメージング研究室 (DBI)  
≒ 理研AIP 脳情報統合解析チーム (川鍋T)
- 計算脳イメージング研究室 (CBI)  
≒ 理研AIP 計算脳ダイナミクスチーム (山下T)

# 理研AIP @ ATR



- 防災科学チーム (上田 修功)
- 脳情報統合解析チーム (川鍋 一晃)
- 計算脳ダイナミクスチーム (山下 宙人)
- iPS細胞連携医学的リスク回避チーム (上田 修功)

理研AIPと京大iPS細胞研の連携ラボ



## メンバー

チームリーダー 上田 修功	研究員 瀬川 一学
技師 永橋 文子	テクニカルスタッフ 松林 由季
テクニカルスタッフ 江浪 青子	テクニカルスタッフ 井上 育代
客員主管研究員 井上 治久	客員研究員 中川 誠人
客員研究員 山本 拓也	

# 現在の関心

## ● 理化学研究所 革新知能統合研究センター (AIP)



- 離散構造・組合せ構造を伴う機械学習
- 幹細胞生物学のための機械学習 (細胞画像 + 深層学習)
- 新しいアルゴリズム・最適化の定式化と求解法
- 透過型電子顕微鏡+機械学習による動的観察

## ● 北海道大学 化学反応創成研究拠点 (WPI-IICReDD)



- 機械学習の実践研究
- 化学反応のデザインと発見のための機械学習
- 分子のグラフ表現の学習と生成
- 量子化学計算 + 機械学習の融合
- 機械発見：探索、実験計画、知識発見

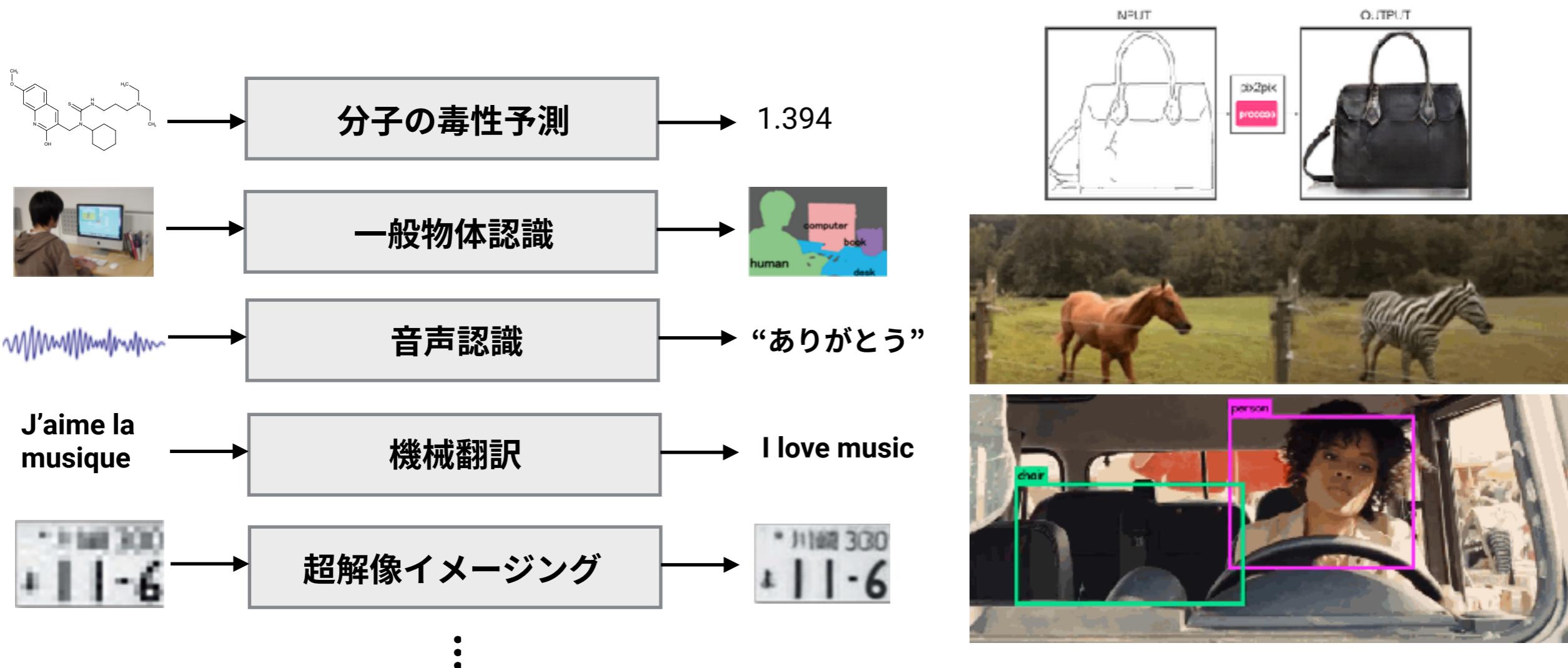
# 今日のテーマ

- **自己紹介 (機械学習と自然科学の境界)**
- **機械学習とは新しいプログラミングの方法**
- **機械学習屋は一体何が楽しいのか？**
  - 分子の表現と機械学習
  - グレイボックス最適化 (演繹 + 帰納)：論理学と統計学の融合？
- **自然科学研究で機械学習を使おうとすると必ずぶつかる本当に難しい問題**
  - データモデリングと予測アルゴリズム (The Two Cultures)
  - 予測か理解か：Rashomon効果, Underspecification, 解釈多様性
  - 人間の認知バイアスに由来する問題：仮説、失敗、成功バイアス、etc.
- **機械学習から機械発見へ**
  - 「発見」は合理化できるのか？さらに自動化できるのか？

# 機械学習とは新手の(難な)プログラミングの方法

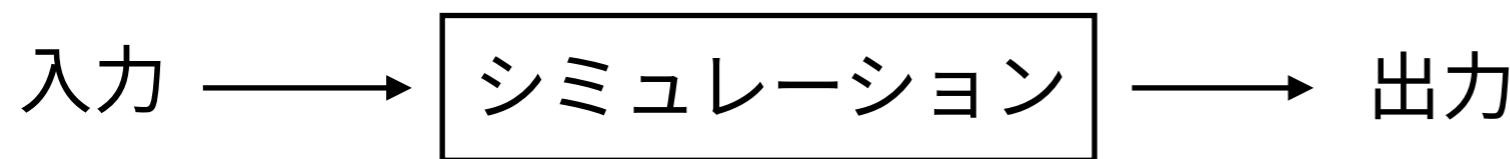
$x \rightarrow \boxed{\text{コンピュータプログラム}} \rightarrow y$

与えられた大量の入出力の見本例を再現できるような入力から出力への変換プログラムを非明示的に生成するための汎用的方法



# 機械学習とは新手の(難な)プログラミングの方法

## 伝統的なプログラミング (演繹的、 rational)



計算のためのロジックはすべて人間が考える

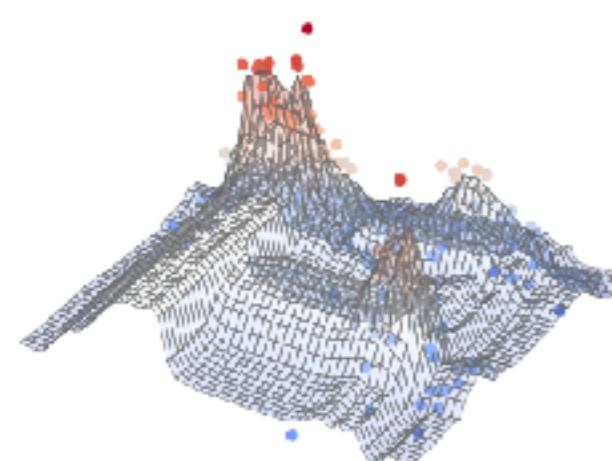
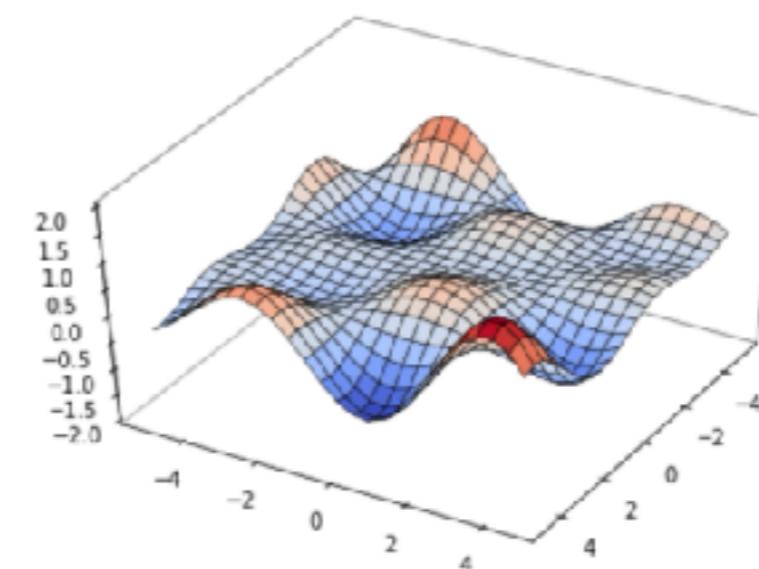
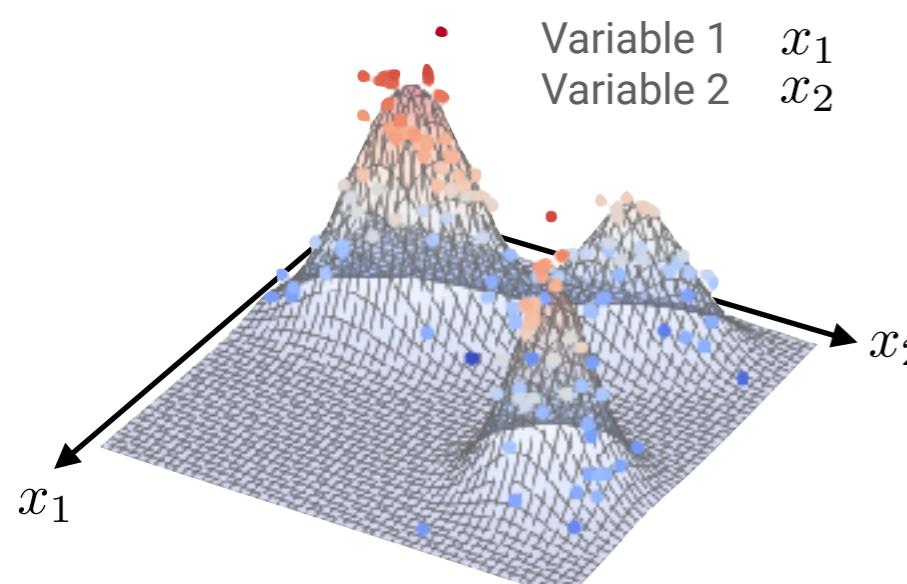
## 新しいプログラミング (帰納的、 empirical)



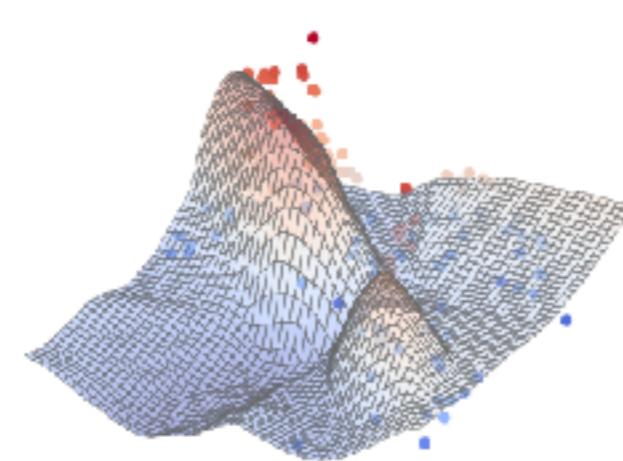
入出力の関係はよく分からないので諦める(!)

↑  
「パラメタで挙動を自由に変えられる汎用形」で  
雛形を用意し、たくさんの入出力の見本例を与えて  
見本例を再現するようパラメタの値を調整する！！

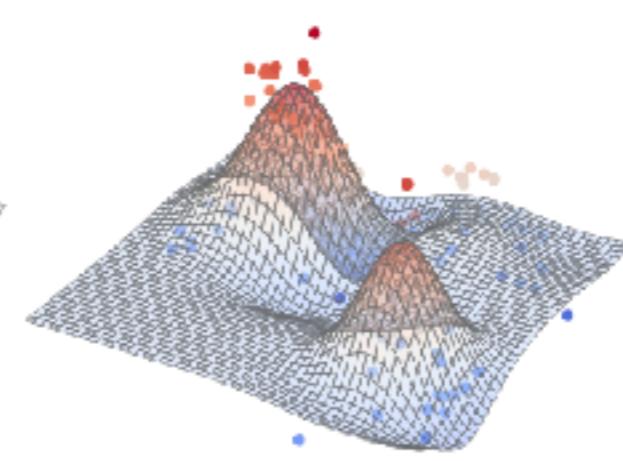
# 機械学習 = 有限の点への関数フィッティング



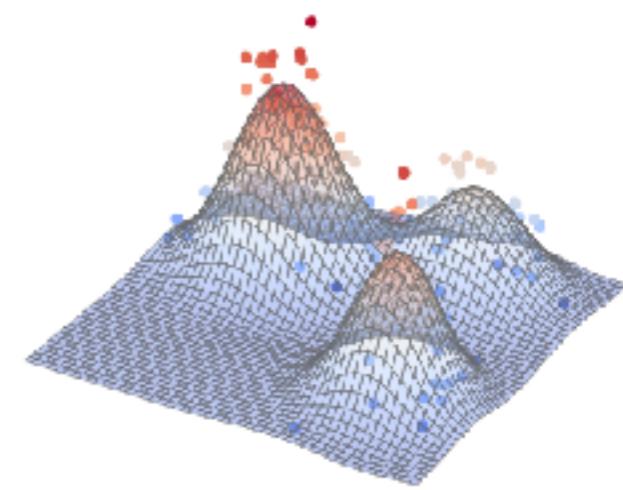
Random Forest



Neural Networks

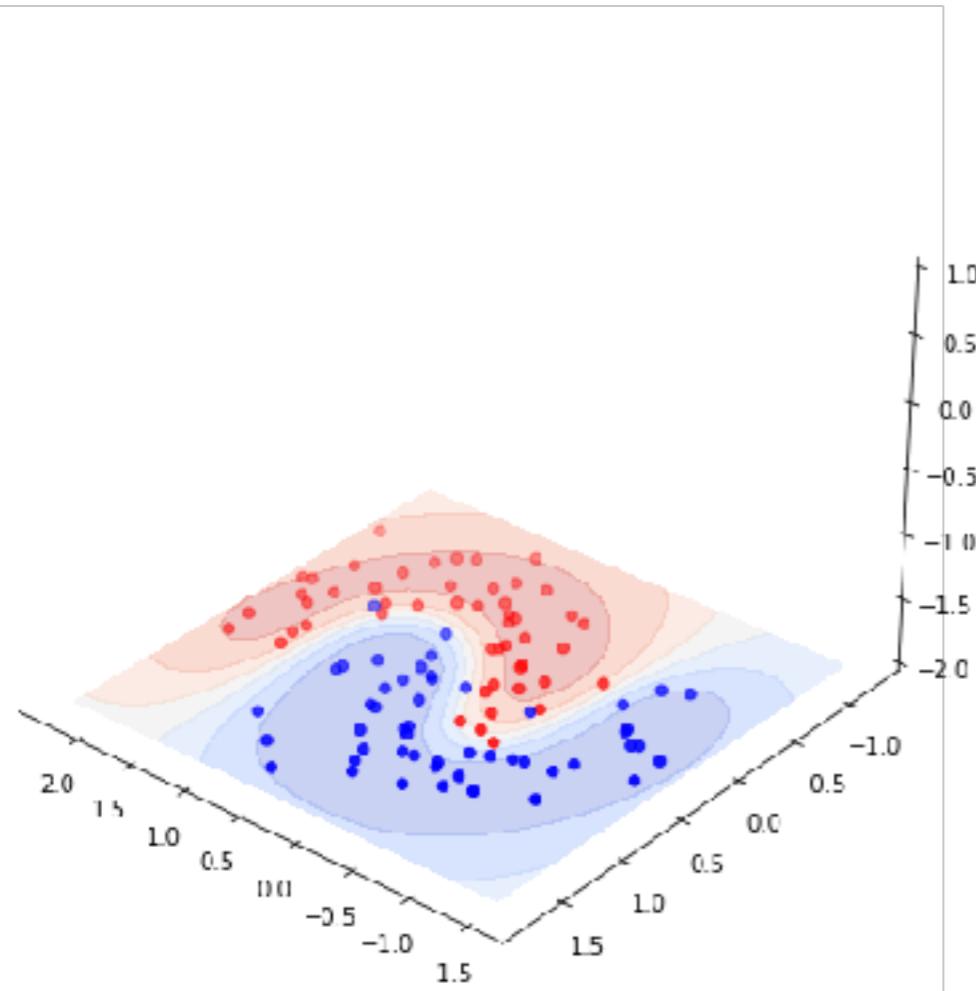


SVR

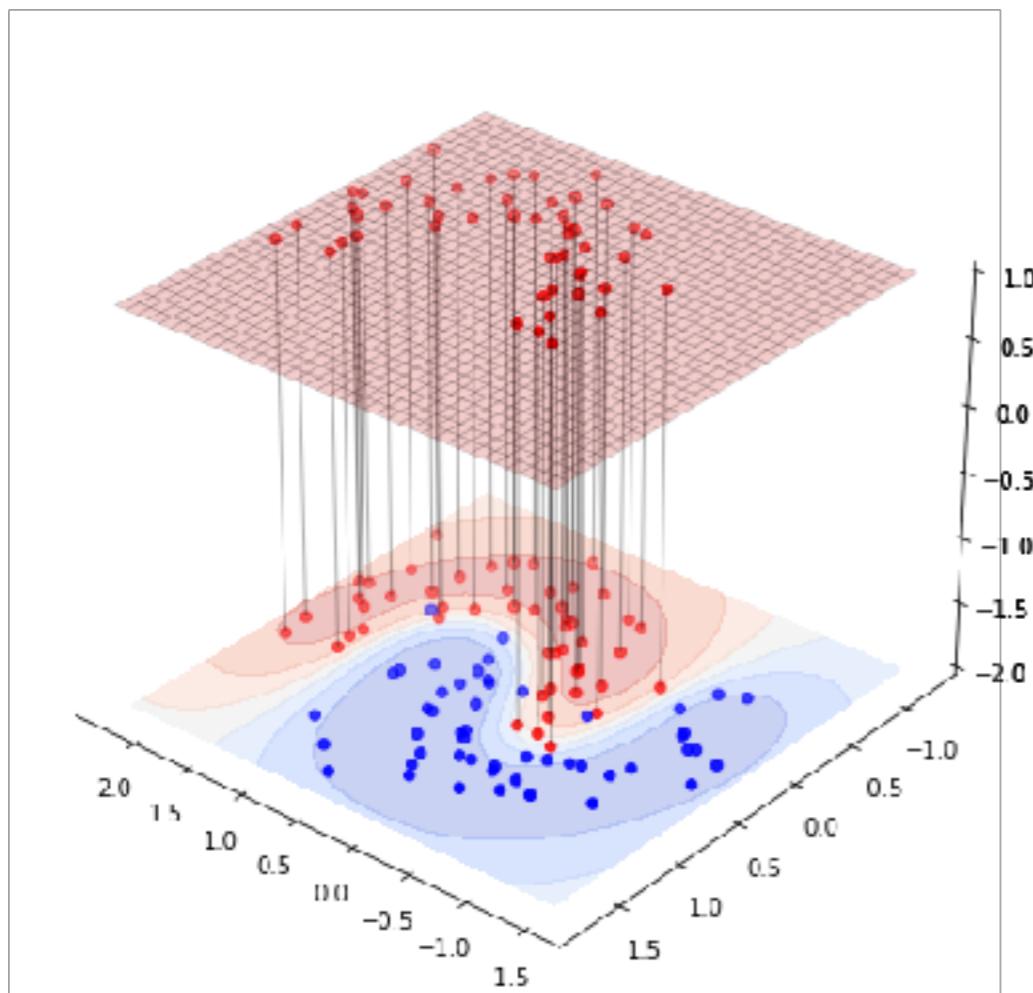


Kernel Ridge

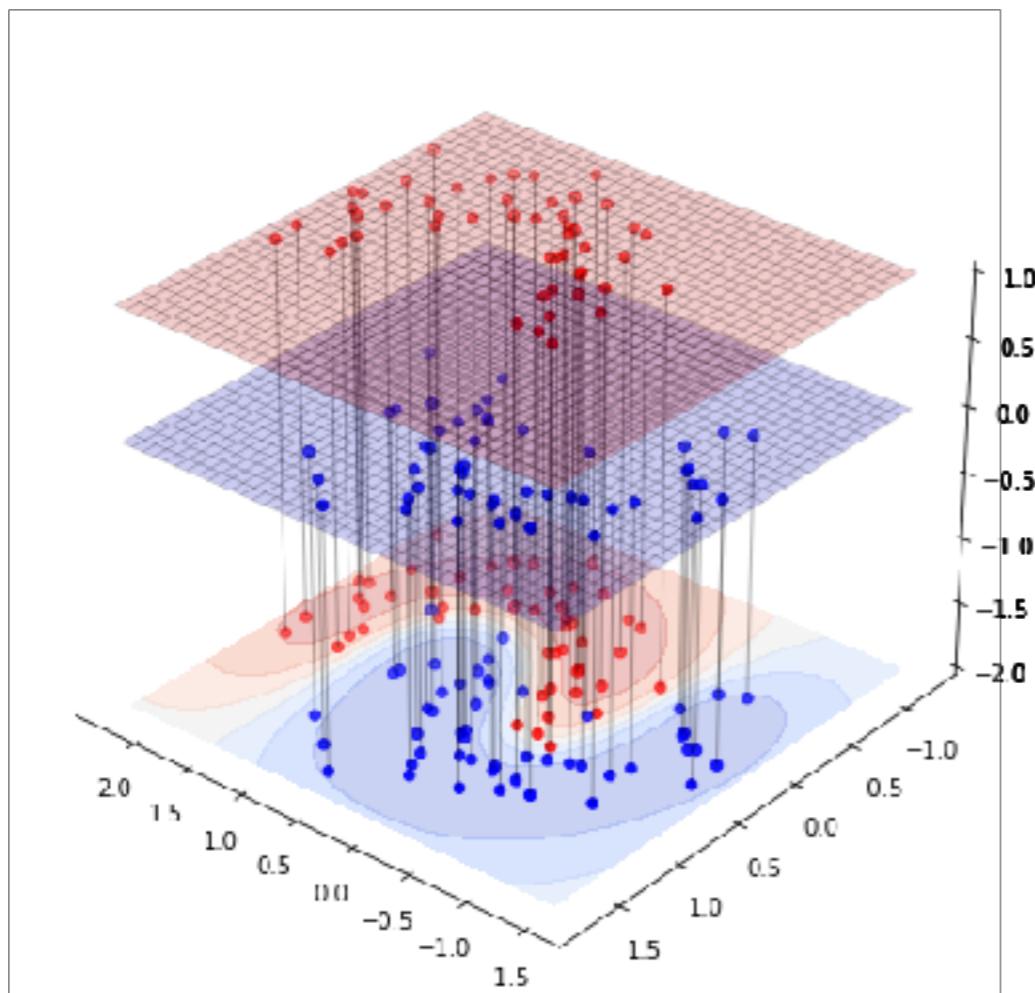
# 分類 (classification) as フィッティング



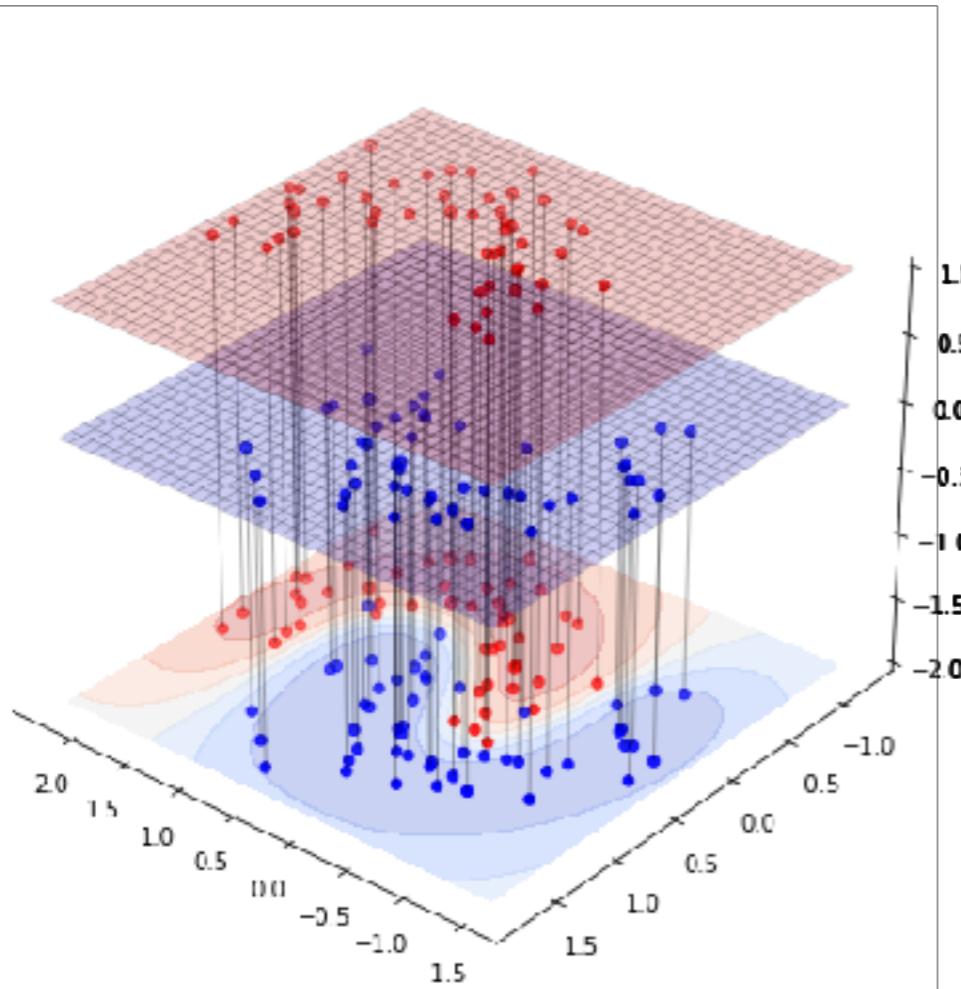
# 分類 (classification) as フィッティング



# 分類 (classification) as フィッティング

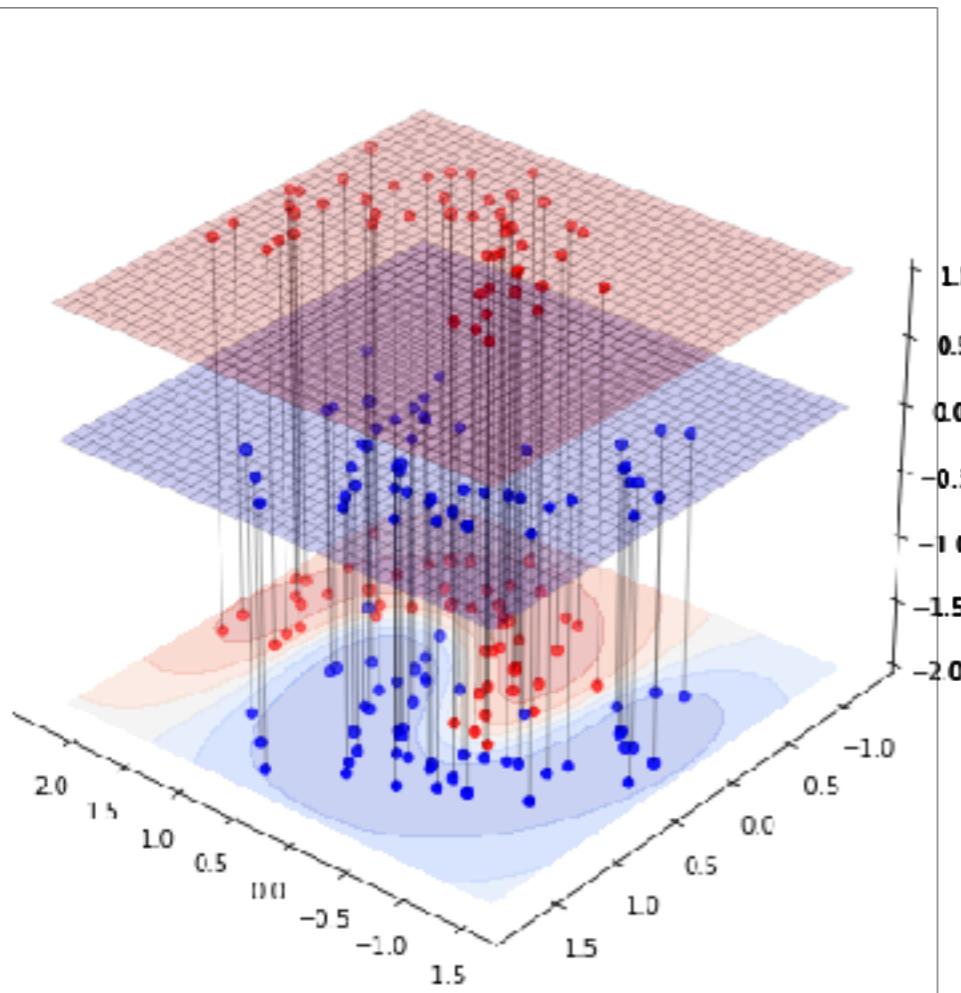


# 分類 (classification) as フィッティング

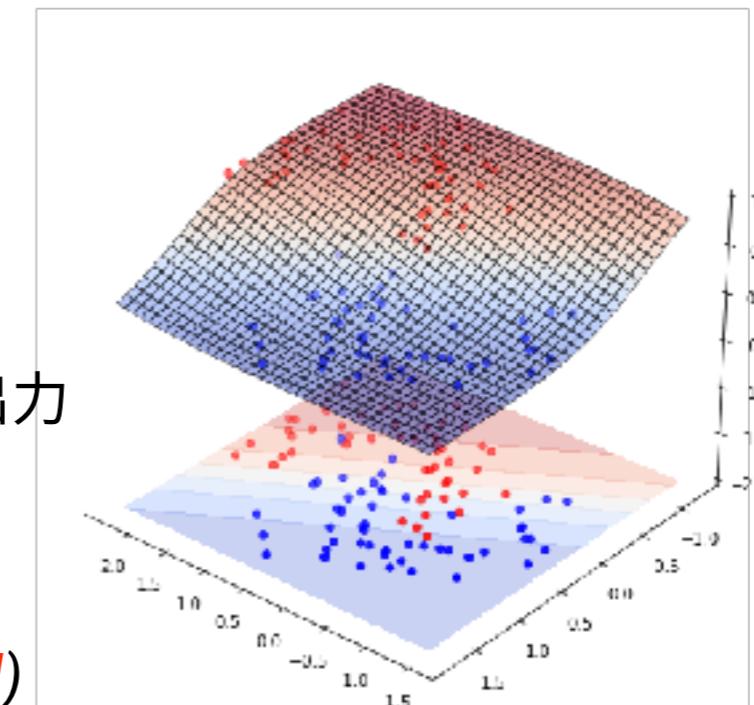


$P(\text{class}=\text{red})$   
0-1の確率値を出力  
(`predict_proba`)  
 $P(\text{class}=\text{blue})$   
 $= 1 - P(\text{class}=\text{red})$

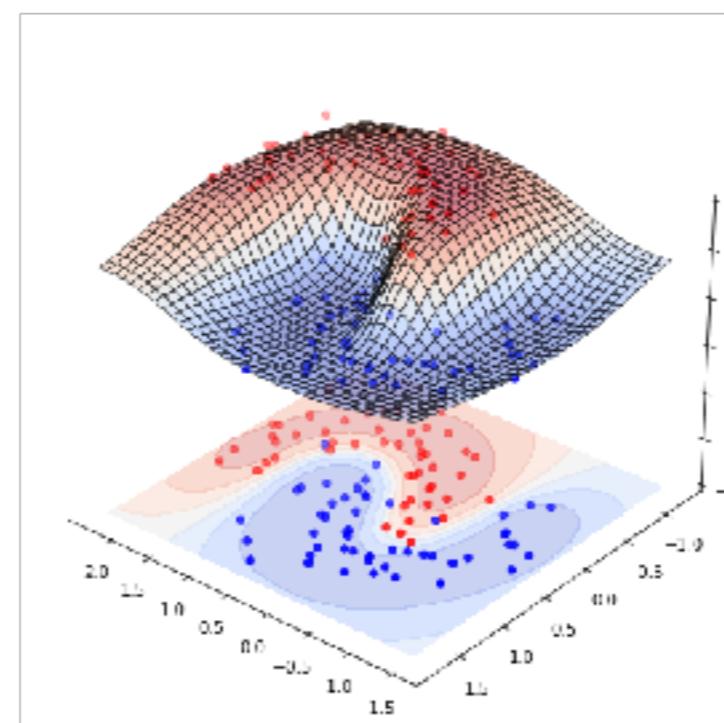
# 分類 (classification) as フィッティング



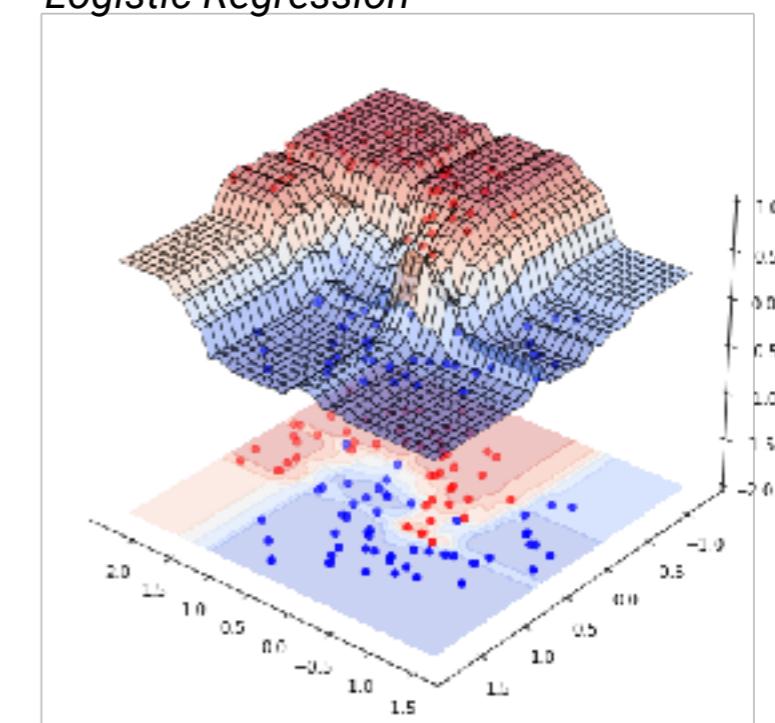
$P(\text{class}=\text{red})$   
0-1の確率値を出力  
(`predict_proba`)  
 $P(\text{class}=\text{blue})$   
 $= 1 - P(\text{class}=\text{red})$



Logistic Regression



Gaussian Process Classifier



Random Forest

# Boring AI (a.k.a. Machine Learning)



Feb 19, 2020, 06:00am EST | 2,187 views

## In Praise Of Boring AI (A.K.A. Machine Learning)

**JC Schutterle** Forbes Councils Member  
**Forbes Technology Council** COUNCIL POST | Paid Program

Matt Velloso, a technical advisor to Microsoft's CEO, got 24,000 likes on this [tweet](#) posted in November 2018: "Difference between machine learning and AI: If it is written in Python, it's probably machine learning. If it is written in PowerPoint, it's probably AI."

:

Whether or not machine learning is paving the way for a sci-fi movie type of AI in the distant future is a pointless question. The benefits of a data-driven approach to automating nitty-gritty processes and transforming organizations as a whole are far from being exhausted. Machine learning offers enough value potential for the new decade. It's time to stop staring at boring PowerPoint decks and start coding in Python. It's time for boring AI.

<https://www.forbes.com/sites/forbestechcouncil/2020/02/19/in-praise-of-boring-ai-a-k-a-machine-learning/>

*The AI frenzy: hope & hype  
"Let's face it:  
So far, the artificial  
intelligence plastered all  
over PowerPoint slides  
hasn't lived up to its hype."*

# Boring AI (a.k.a. Machine Learning)



Feb 19, 2020, 06:00am EST | 2,187 views

## In Praise Of Boring AI (A.K.A. Machine Learning)

**JC Schutterle** Forbes Councils Member  
**Forbes Technology Council** COUNCIL POST | Paid Program

Matt Velloso, a technical advisor to Microsoft's CEO, got 24,000 likes on this [tweet](#) posted in November 2018: "Difference between machine learning and AI: If it is written in Python, it's probably machine learning. If it is written in PowerPoint, it's probably AI."

:

Whether or not machine learning is paving the way for a sci-fi movie type of AI in the distant future is a pointless question. The benefits of a data-driven approach to automating nitty-gritty processes and transforming organizations as a whole are far from being exhausted. Machine learning offers enough value potential for the new decade. It's time to stop staring at boring PowerPoint decks and start coding in Python. It's time for boring AI.

<https://www.forbes.com/sites/forbestechcouncil/2020/02/19/in-praise-of-boring-ai-a-k-a-machine-learning/>

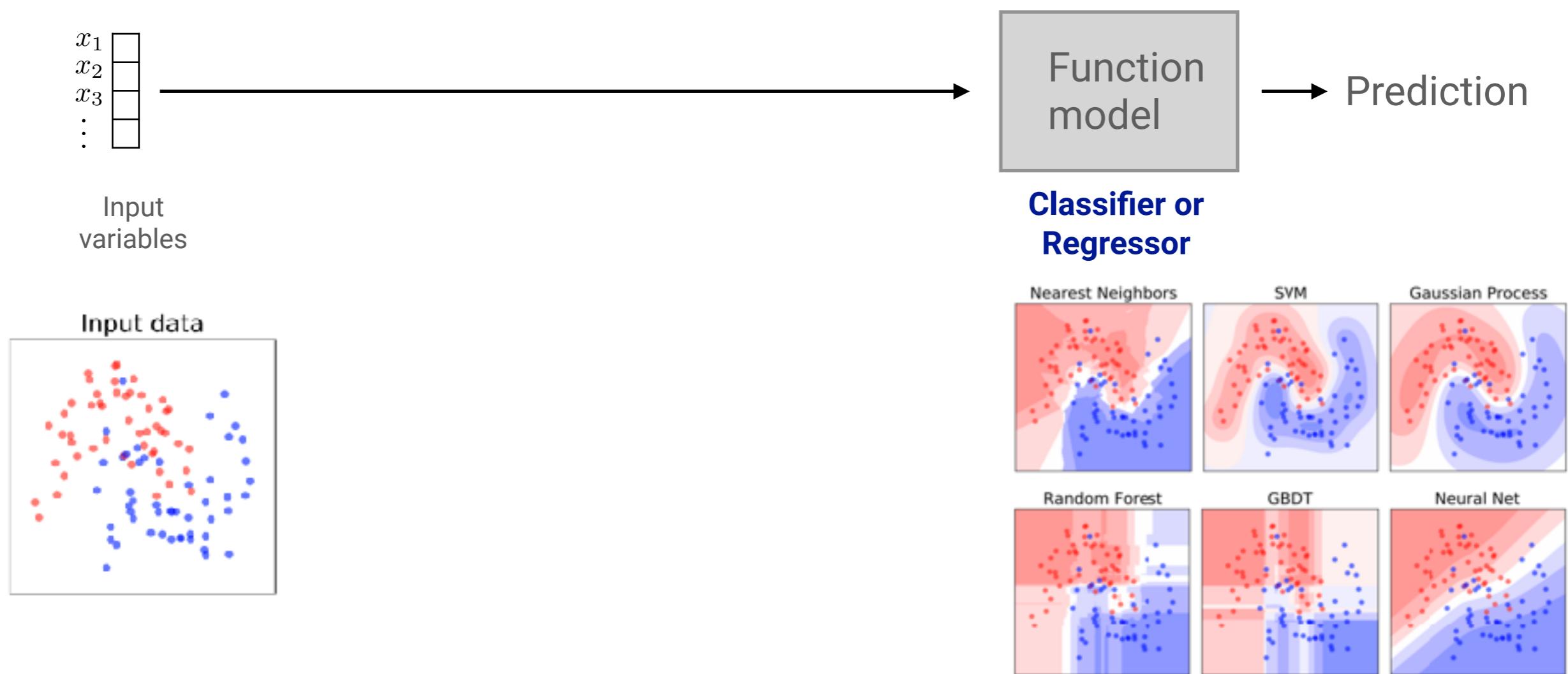
*The AI frenzy: hope & hype*  
*"Let's face it:*  
*So far, the artificial*  
*intelligence plastered all*  
*over PowerPoint slides*  
*hasn't lived up to its hype."*



*From AAAI-20 Oxford-Style Debate*

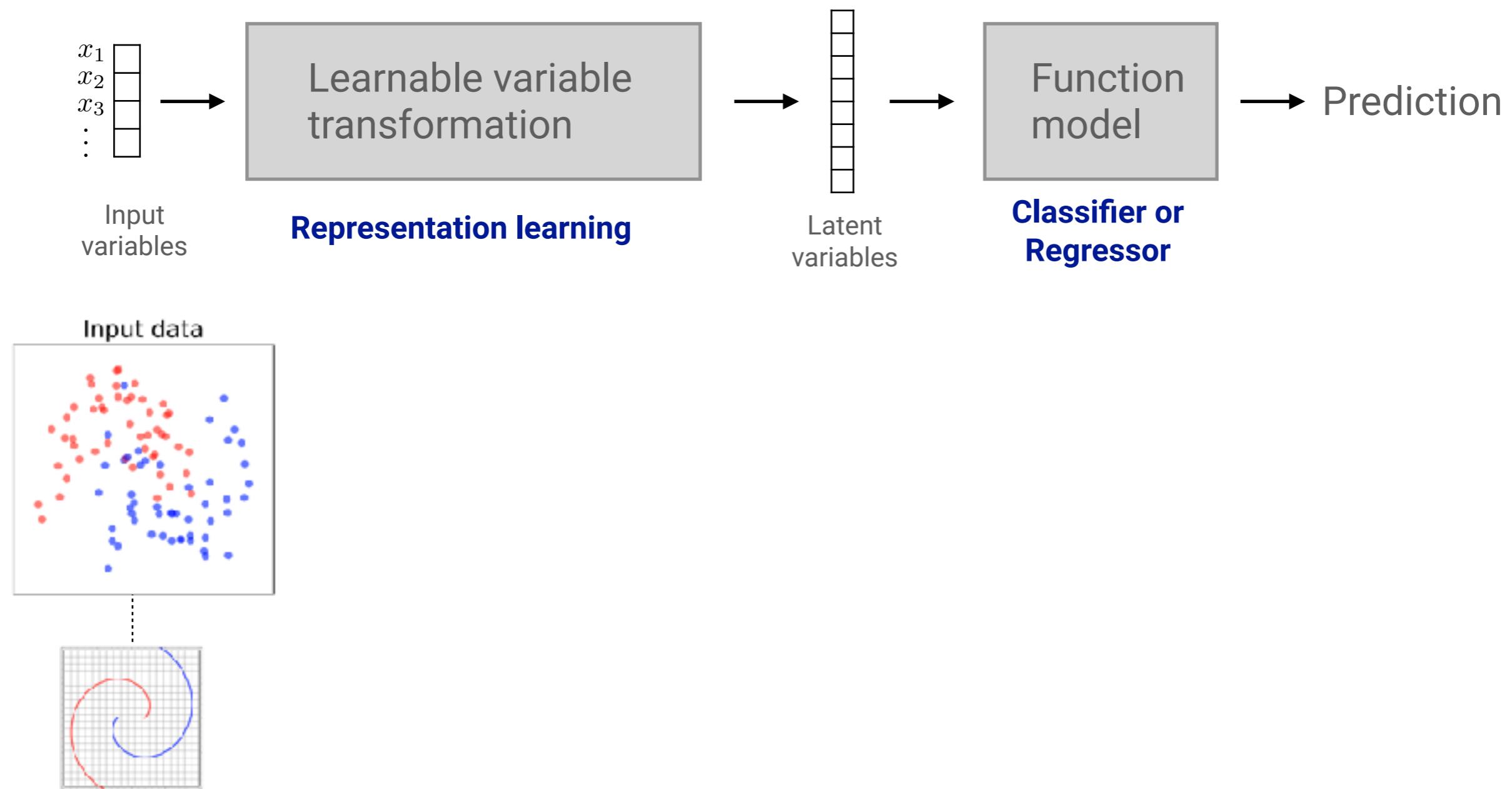
# 関数フィッティングとしての機械学習

## 標準的な機械学習モデル



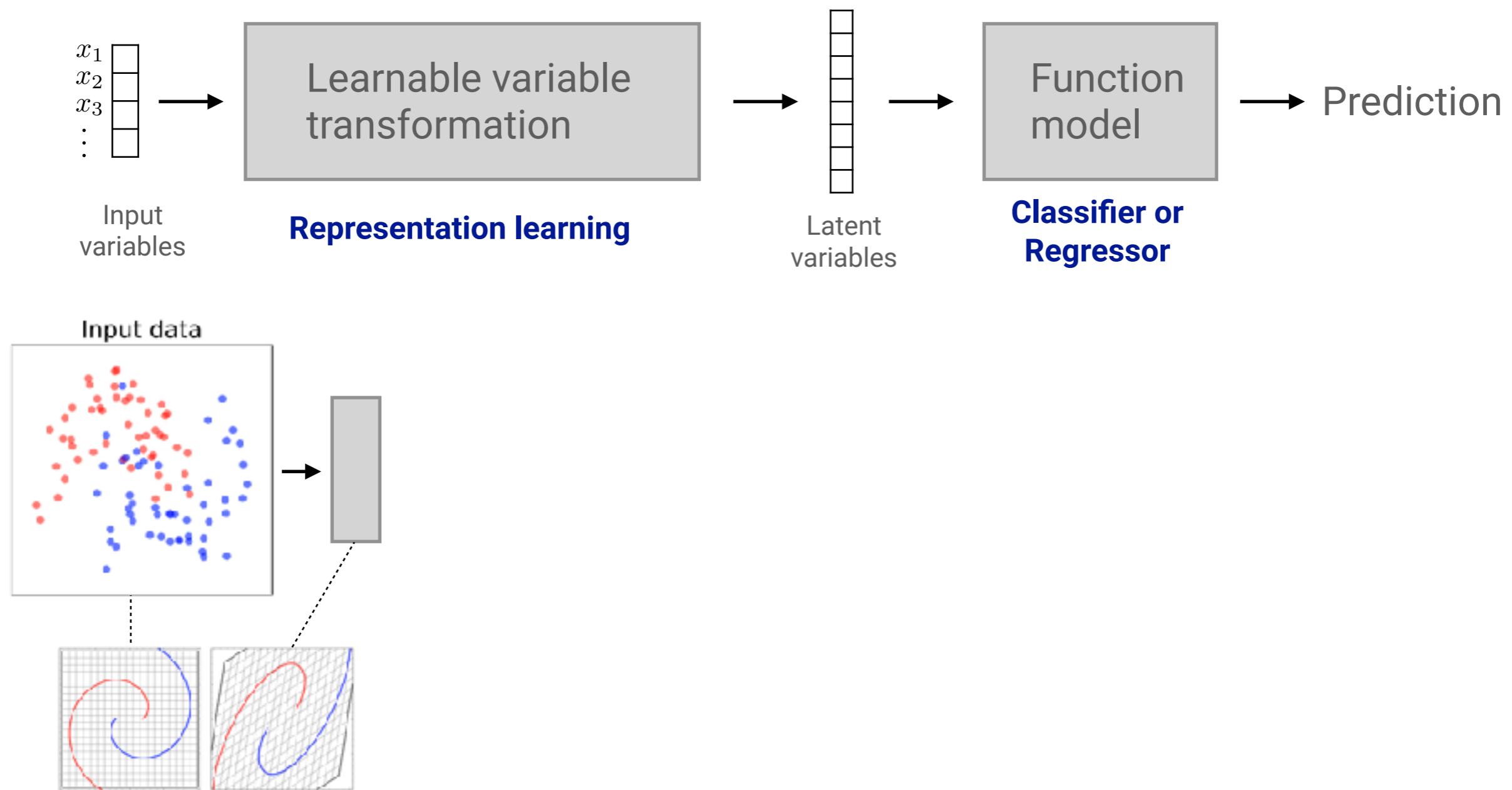
# 表現学習（良い潜在特徴量のデータからの抽出）

最近の深層学習では回帰・分類の前に良い潜在変数表現への変換を行う！  
(2つのブロックの合成を関数フィッティングとして一気通貫で最適化する)



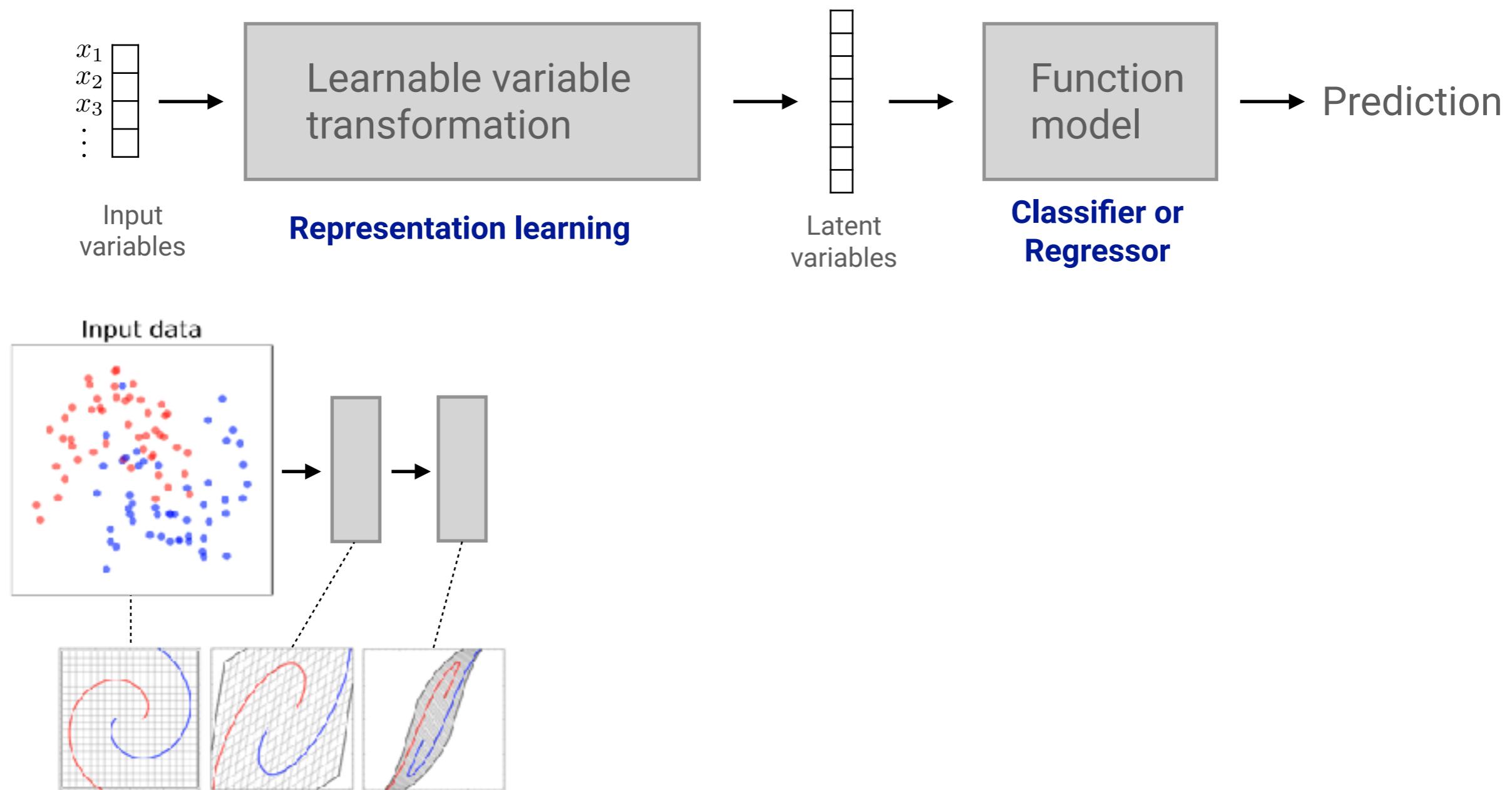
# 表現学習 (良い潜在特徴量のデータからの抽出)

最近の深層学習では回帰・分類の前に**良い潜在変数表現への変換**を行う！  
 (2つのブロックの合成を関数フィッティングとして一気通貫で最適化する)



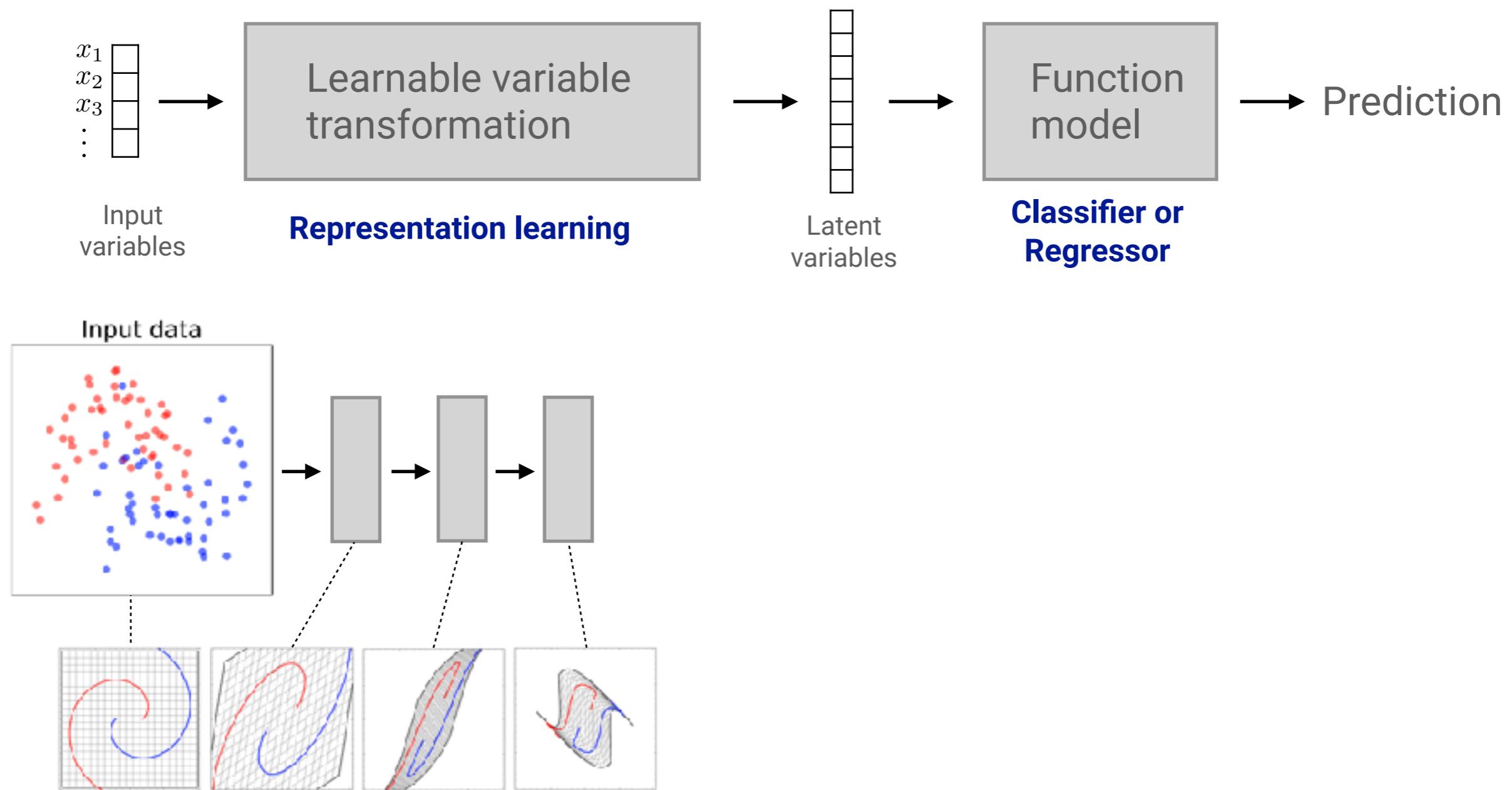
# 表現学習 (良い潜在特徴量のデータからの抽出)

最近の深層学習では回帰・分類の前に**良い潜在変数表現への変換**を行う！  
 (2つのブロックの合成を関数フィッティングとして一気通貫で最適化する)



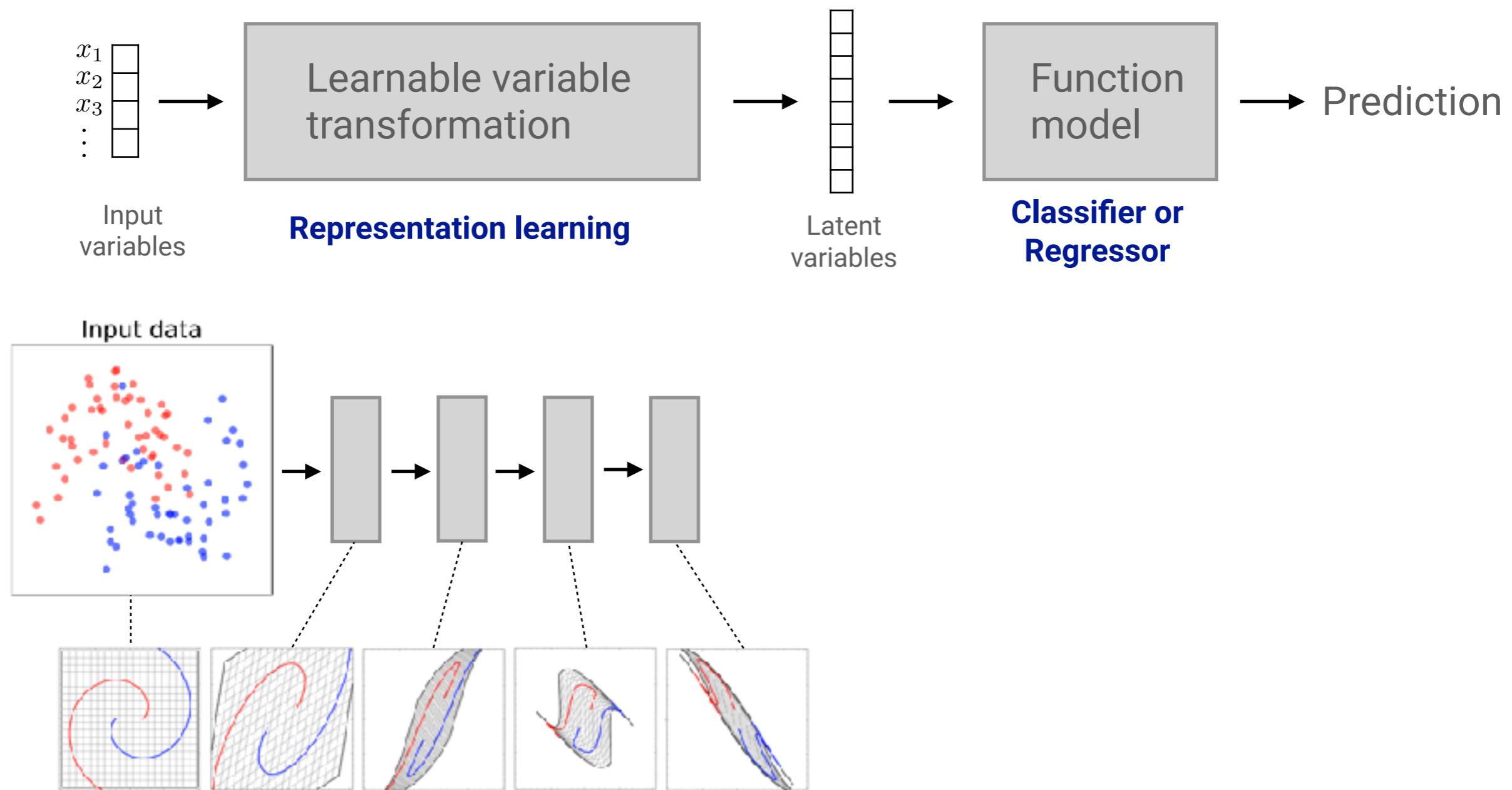
# 表現学習 (良い潜在特徴量のデータからの抽出)

最近の深層学習では回帰・分類の前に**良い潜在変数表現への変換**を行う！  
 (2つのブロックの合成を関数フィッティングとして一気通貫で最適化する)



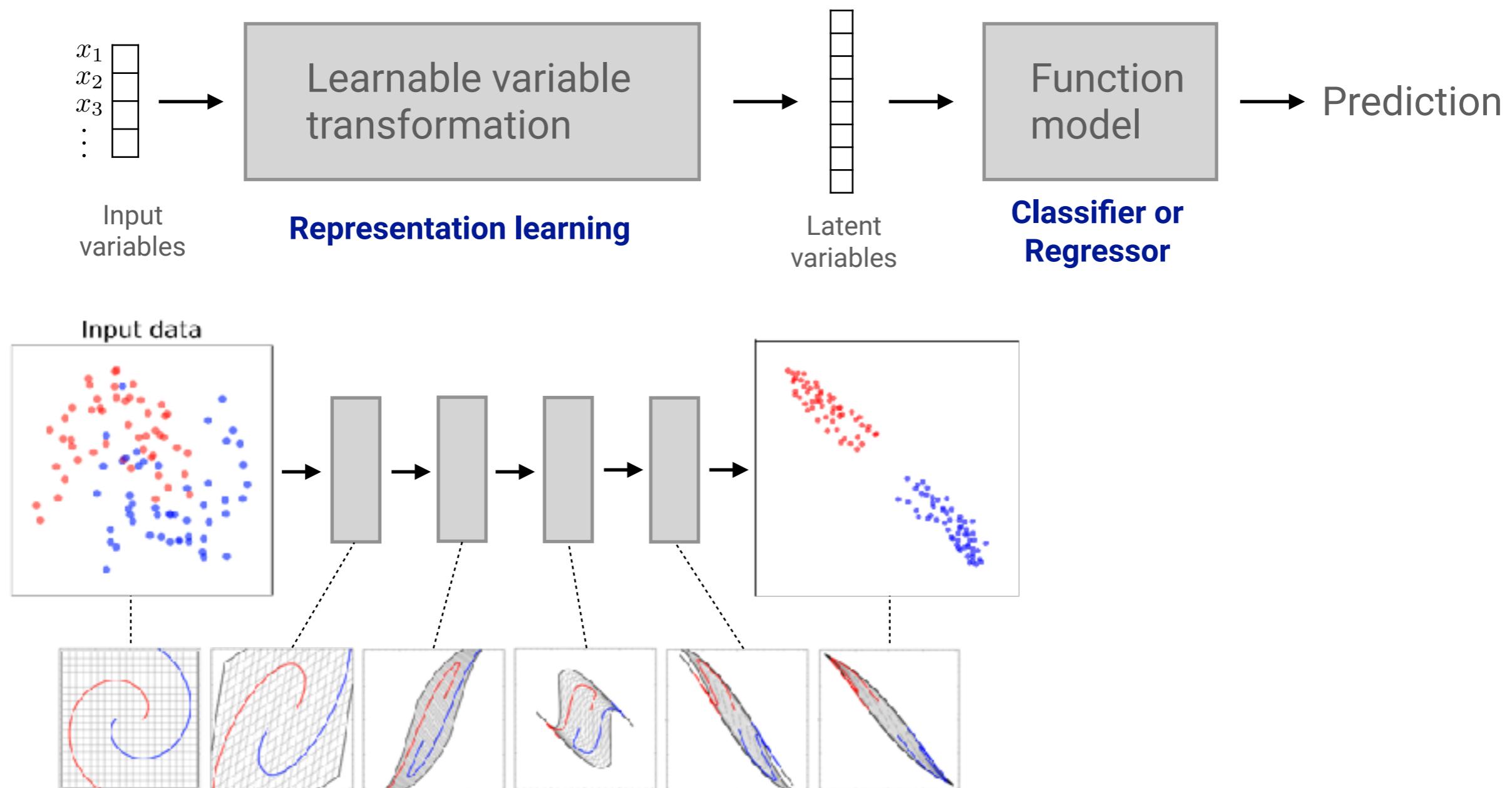
# 表現学習 (良い潜在特徴量のデータからの抽出)

最近の深層学習では回帰・分類の前に**良い潜在変数表現への変換**を行う！  
 (2つのブロックの合成を関数フィッティングとして一気通貫で最適化する)



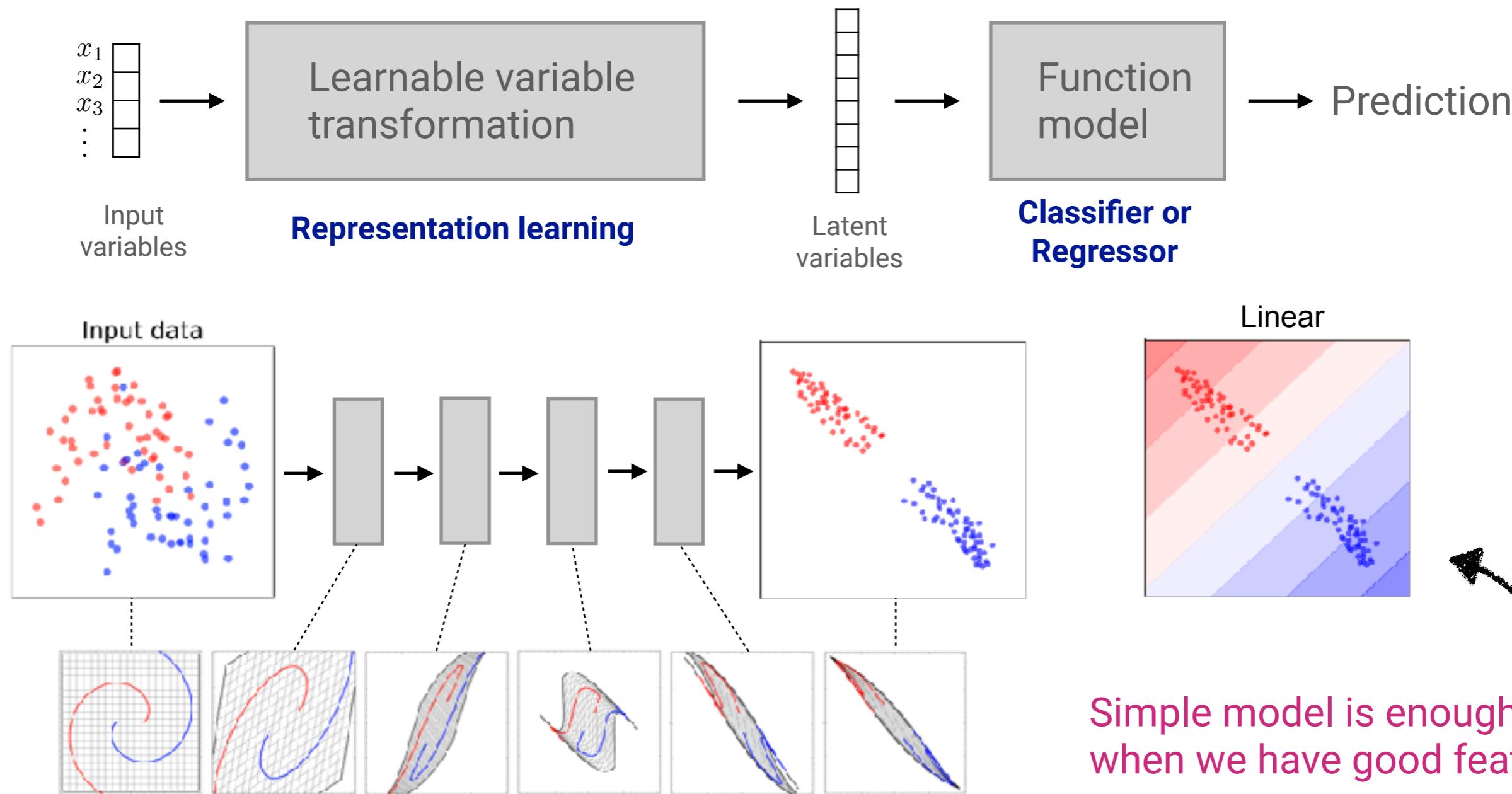
# 表現学習 (良い潜在特徴量のデータからの抽出)

最近の深層学習では回帰・分類の前に**良い潜在変数表現への変換**を行う！  
 (2つのブロックの合成を関数フィッティングとして一気通貫で最適化する)

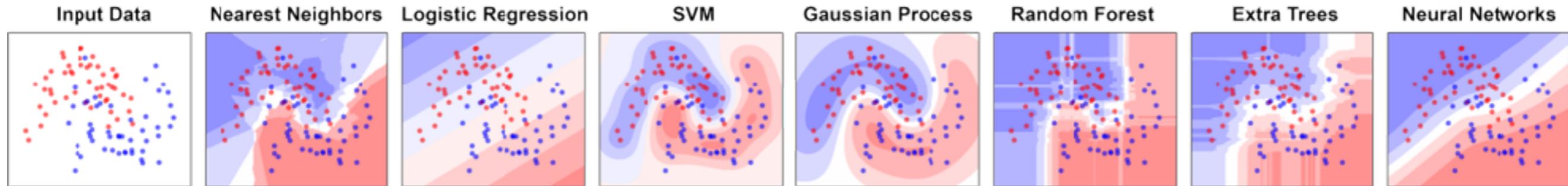


# 表現学習 (良い潜在特徴量のデータからの抽出)

最近の深層学習では回帰・分類の前に**良い潜在変数表現への変換**を行う！  
 (2つのブロックの合成を関数フィッティングとして一気通貫で最適化する)

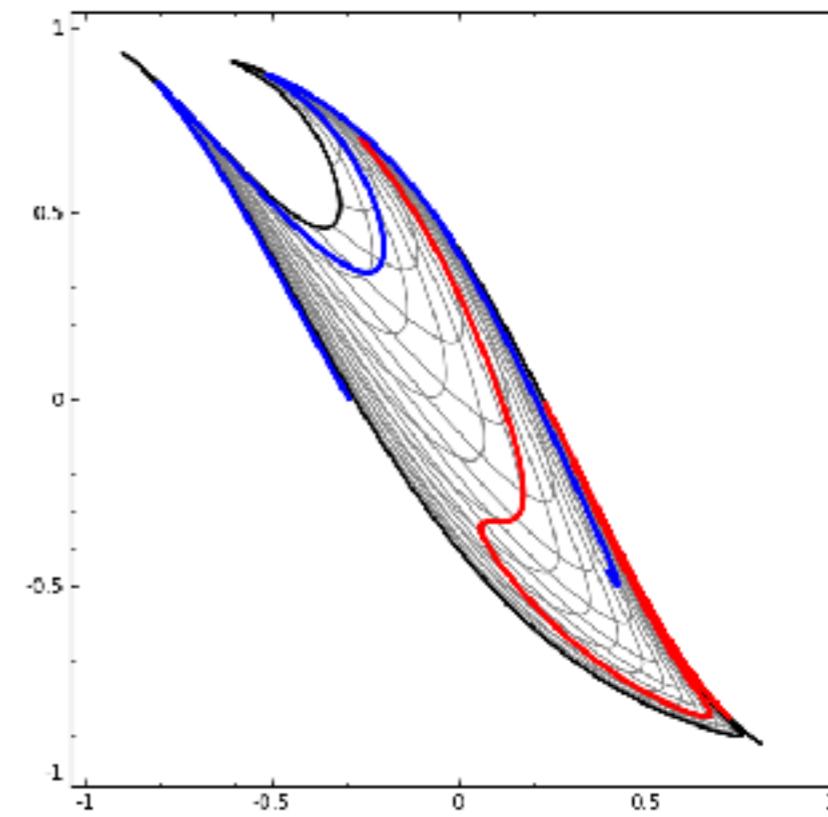
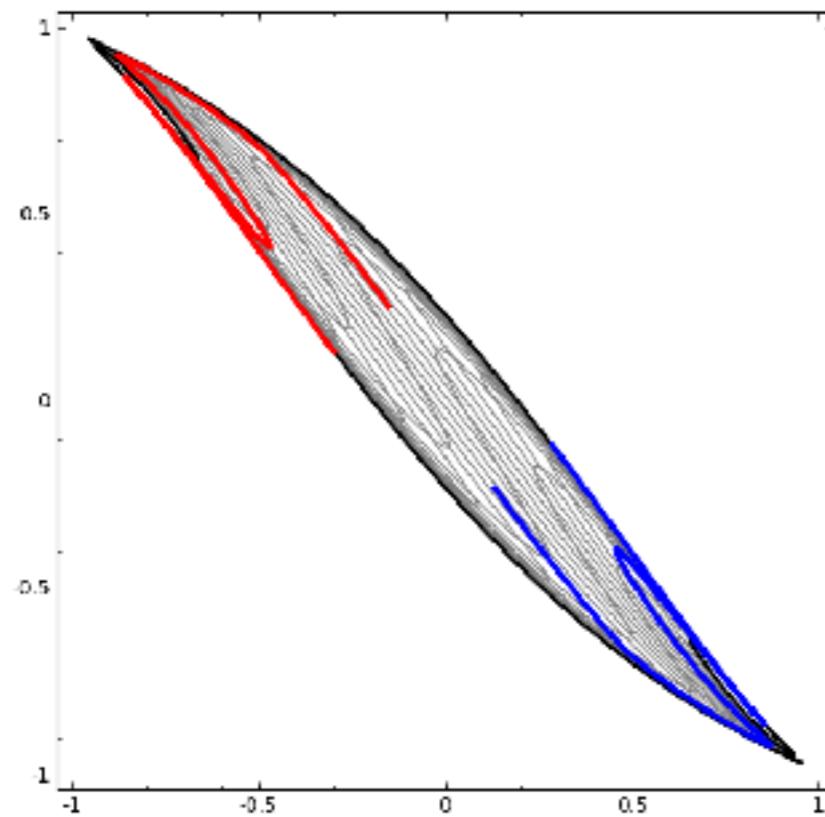


# 表現学習 (良い潜在特徴量のデータからの抽出)



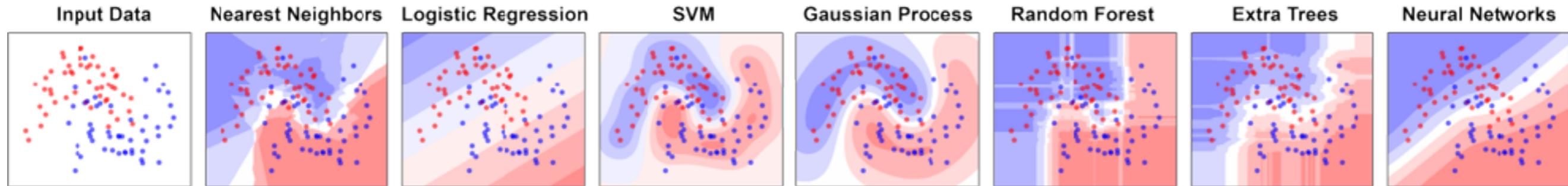
線形分離可能な表現への変換を学習

このタスクは実践的には依然むずかしく  
間違えると元より酷くなりうる..



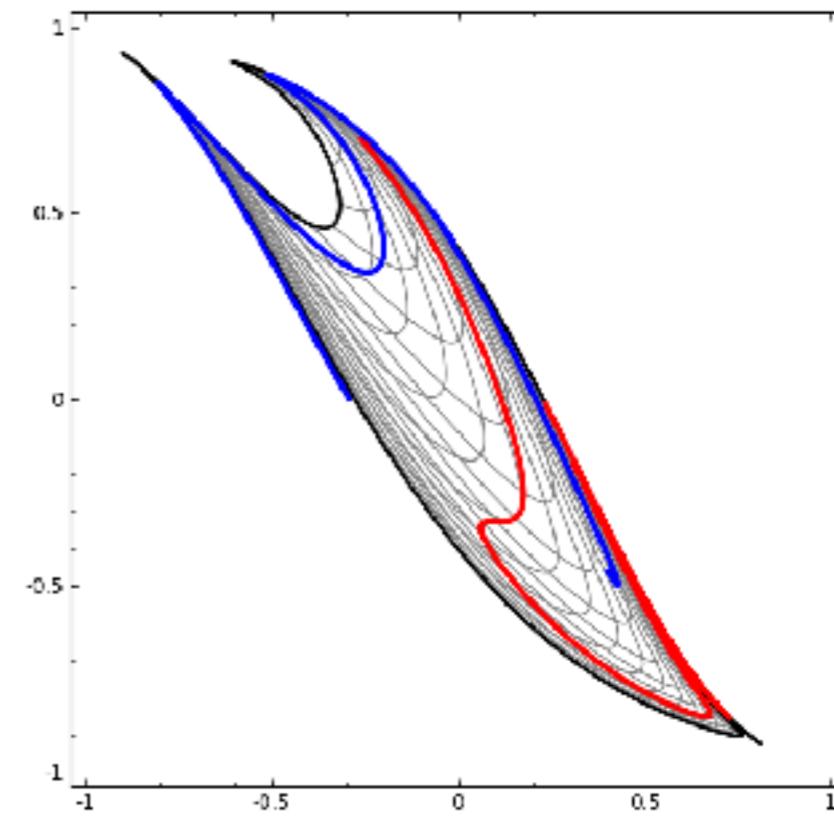
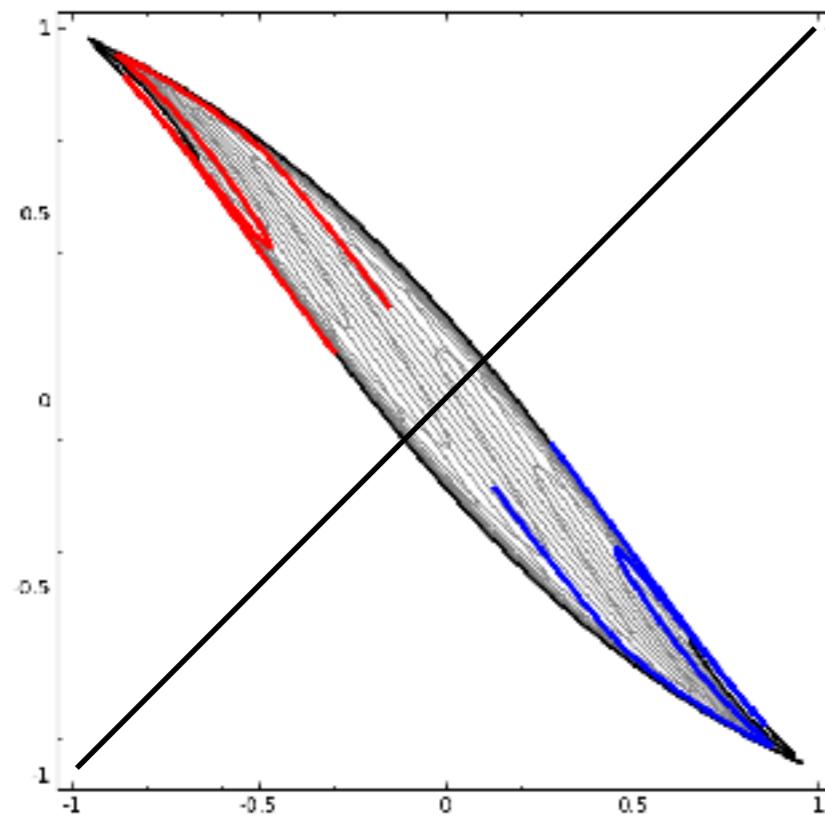
<https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>

# 表現学習 (良い潜在特徴量のデータからの抽出)



線形分離可能な表現への変換を学習

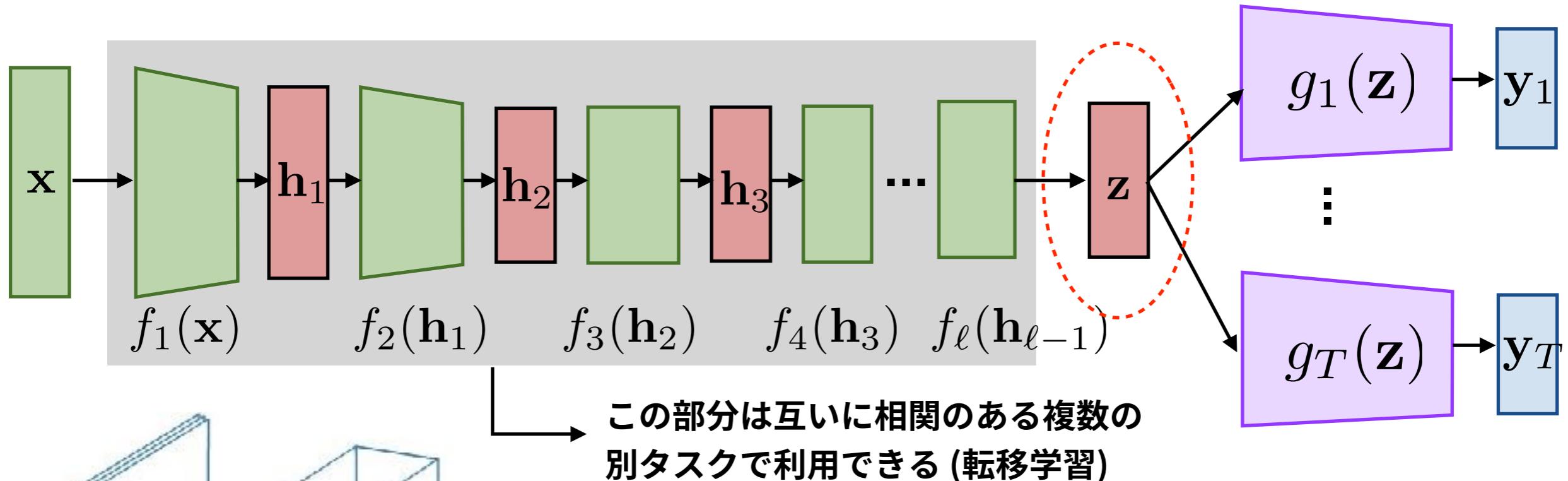
このタスクは実践的には依然むずかしく  
間違えると元より酷くなりうる..



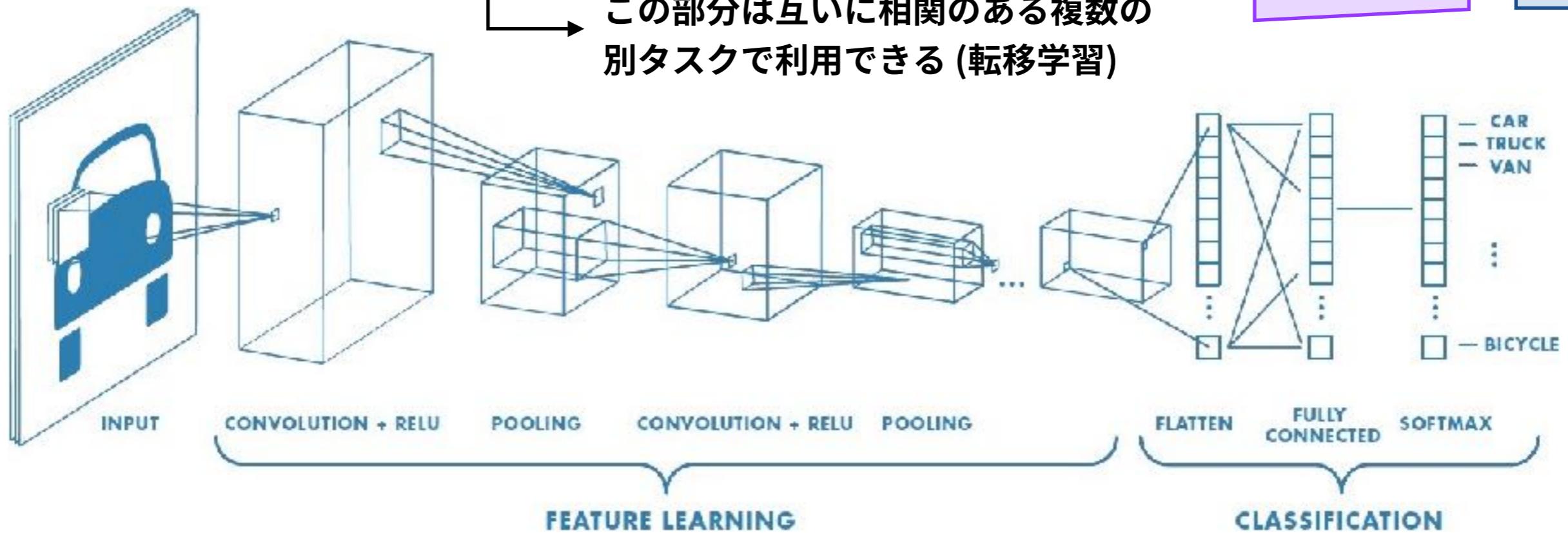
<https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>

# 表現学習(良い潜在特徴量)の別の下流タスクへの転移

良い潜在変数表現への変換をデータから学習

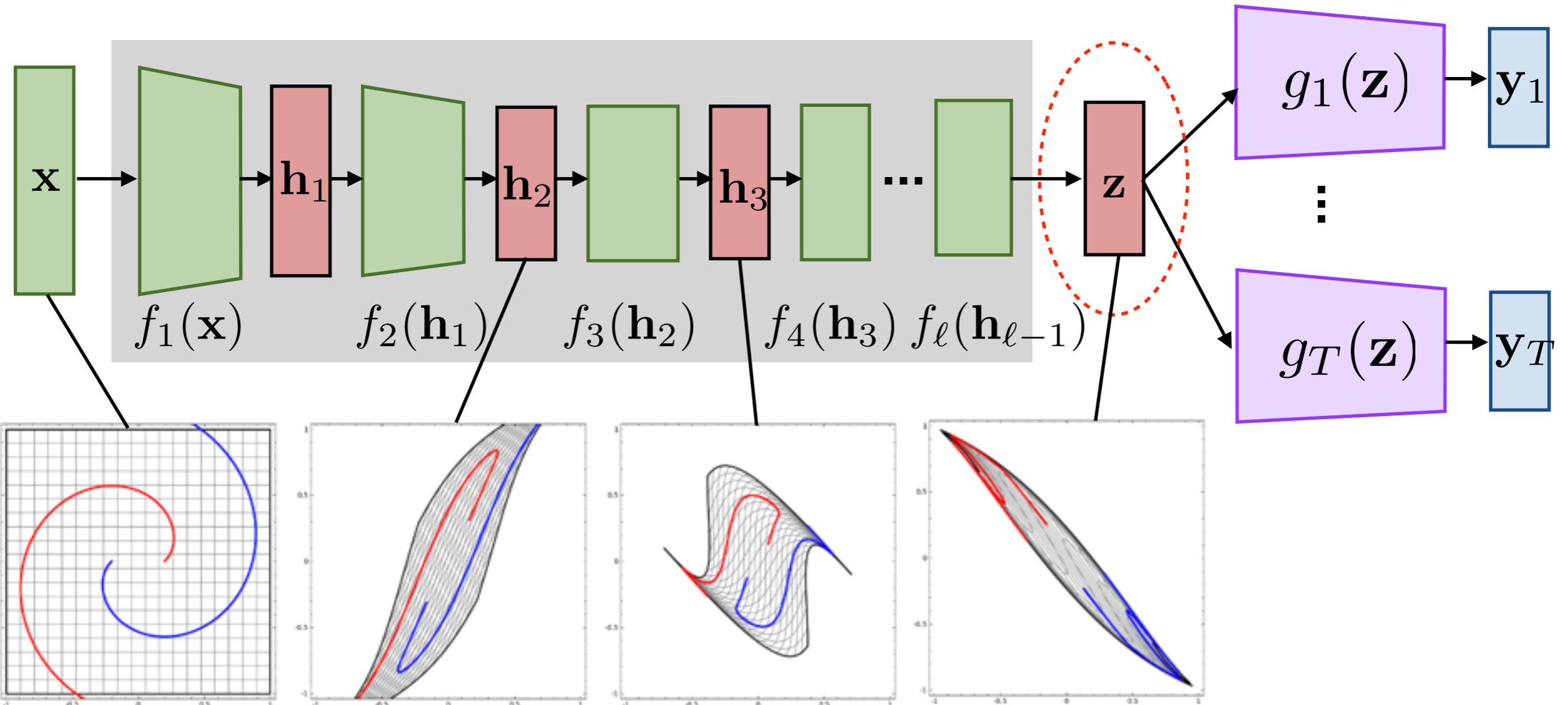


タスクごとに回帰・分類



# 表現学習(良い潜在特徴量)の別の下流タスクへの転移

良い潜在変数表現への変換をデータから学習



タスクごとに回帰・分類

- ✓ データ点を入力表現ではなく **潜在変数表現**において内挿することになる。
- ✓ 共通の「良い潜在変数表現」を持つタスクで「表現学習ブロック」だけを学習できる可能性を持つ。(大規模データでの事前学習→小規模例へ転移学習)

# 機械学習の現代的な側面

- モデルパラメタ数がとんでもなく多い！

現代の機械学習は1,750億個のパラメタ(自由度)を持つモデルを数十万の次元を持つ数千万個のデータにフィッティングしていく直感が効かない非自明な状況！

ResNet50: 26 million params

ResNet101: 45 million params

EfficientNet-B7: 66 million params

VGG19: 144 million params

12-layer, 12-heads BERT: 110 million params

24-layer, 16-heads BERT: 336 million params

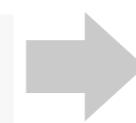
GPT-2 XL: 1558 million params

GPT-3: 175 billion params

- 自動微分系の発展によりプログラムとして書ければ何でも機械学習可能に

```
import torch
from torch.nn.parameter import Parameter

class func():
    def __init__(self):
        self.p1 = Parameter(torch.rand(1), requires_grad=True)
        self.p2 = Parameter(torch.rand(1), requires_grad=True)
    def calc(self, x):
        if x.min() > 0.5:
            return torch.sum(self.p1*torch.sin(x) + self.p2 * x**2 + x.min())
        else:
            return torch.mean(self.p1*torch.cosh(x) + (self.p2 + self.p1)/x.max())
```



```
prog = func()
v = prog.calc(torch.rand(10))

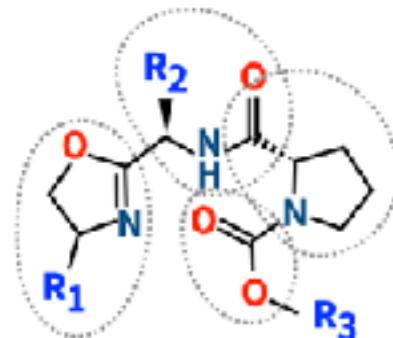
v
tensor(3.0217, grad_fn=<MeanBackward0>)

v.backward()
prog.p1.grad, prog.p2.grad
(tensor([2.2671]), tensor([1.0052]))
```

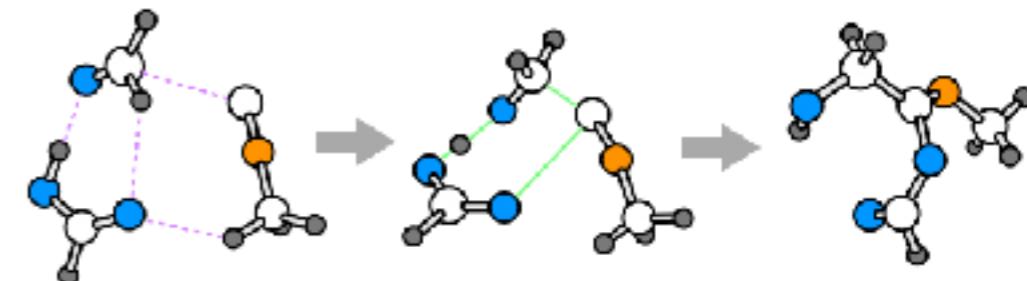
# 今日のテーマ

- **自己紹介 (機械学習と自然科学の境界)**
- **機械学習とは新しいプログラミングの方法**
- **機械学習屋は一体何が楽しいのか？**
  - 分子の表現と機械学習
  - グレイボックス最適化 (演繹 + 帰納)：論理学と統計学の融合？
- **自然科学研究で機械学習を使おうとすると必ずぶつかる本当に難しい問題**
  - データモデリングと予測アルゴリズム (The Two Cultures)
  - 予測か理解か：Rashomon効果, Underspecification, 解釈多様性
  - 人間の認知バイアスに由来する問題：仮説、失敗、成功バイアス、etc.
- **機械学習から機械発見へ**
  - 「発見」は合理化できるのか？さらに自動化できるのか？

# 分子は「組合せ的」な側面をもつ



$R_1$	$R_2$	$R_3$
H	Methyl	Ethyl
Hydrogen		
Phenyl	Benzyl	Isopropyl
Carboxyl	Cyclohexyl	Tert-butyl
		Trifluoromethyl
		Solvomethyl
		...



c&en

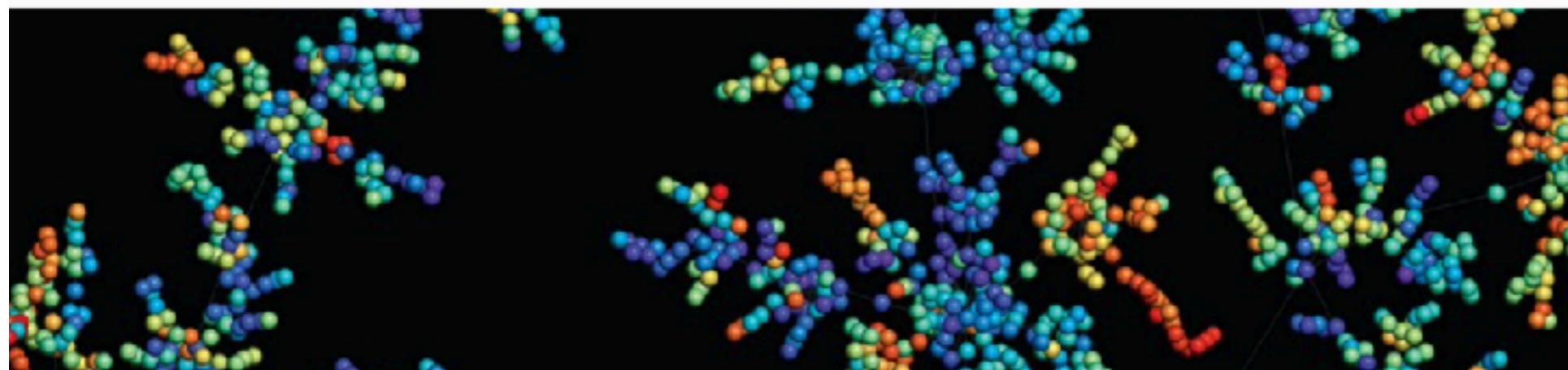
COMPUTATIONAL CHEMISTRY

## Exploring chemical space: Can AI take us where no human has gone before?

Artificial intelligence is helping us find novel, useful molecules. For the field to really take off, though, these tools will need to be accessible to the wider chemistry community

by Sam Lemonick

April 6, 2020 | A version of this story appeared in **Volume 98, Issue 13**



BY THE NUMBERS

$10^{180}$

An upper estimate of the number of possible molecules

$10^{80}$

Estimated number of atoms in the universe

$10^{60}$

An estimate of the number of possible small organic molecules

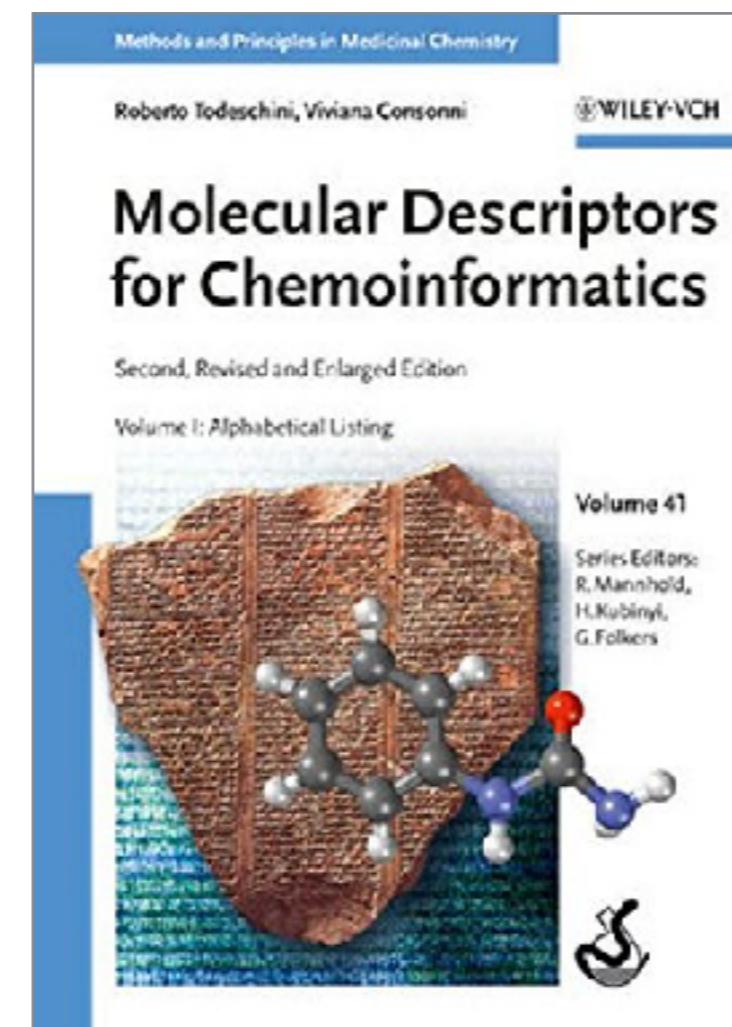
$10^8$

The number of organic and inorganic substances in the CAS database

# 分子記述子

- 実験的な計測量
- 計算的な記述子
  - 0D 記述子
    - constitutional descriptors
    - count descriptors
  - 1D 記述子
    - list of structural fragments
    - fingerprints
  - 2D 記述子
    - graph invariants
  - 3D 記述子
    - 3D MoRSE, WHIM, GETAWAY, ...
    - quantum-chemical descriptors
    - size, steric, surface, volume, etc.
  - 4D 記述子
    - GRID, CoMFA, Volsurf, ...

...more than 3,300 descriptors



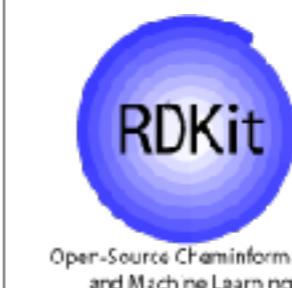
Todeschini and Consonni, *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, 2009.  
<https://doi.org/10.1002/9783527628766>

5,270 descriptors

**DRAGON 7.0**



商用の記述子ソフトウェア



`rdkit.Chem`

- Descriptors
- Descriptors3D
- GraphDescriptors
- Fingerprints
- ChemicalFeatures
- ChemicalForceFields

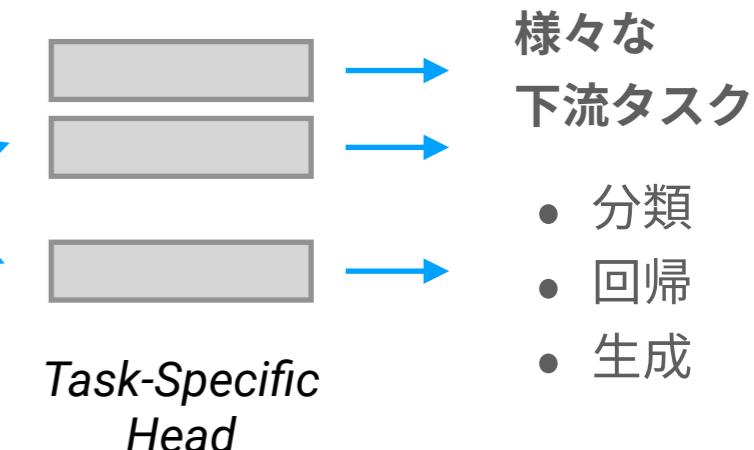
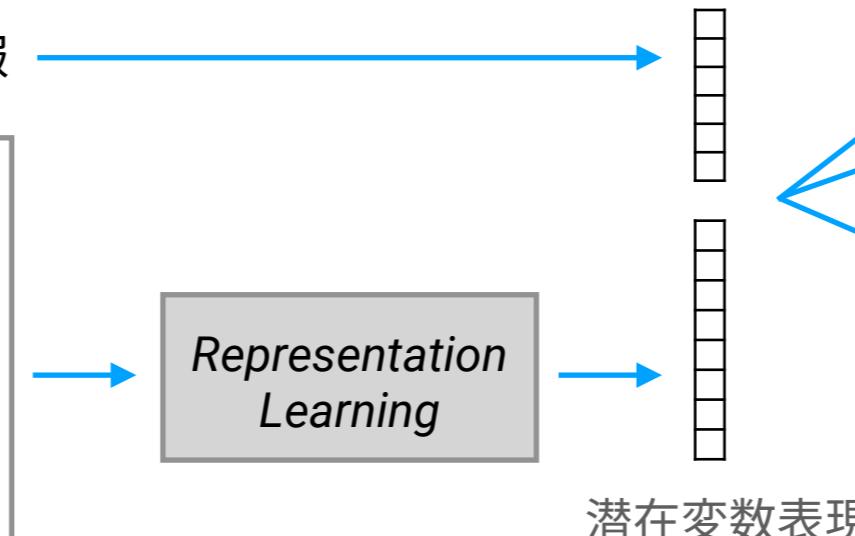
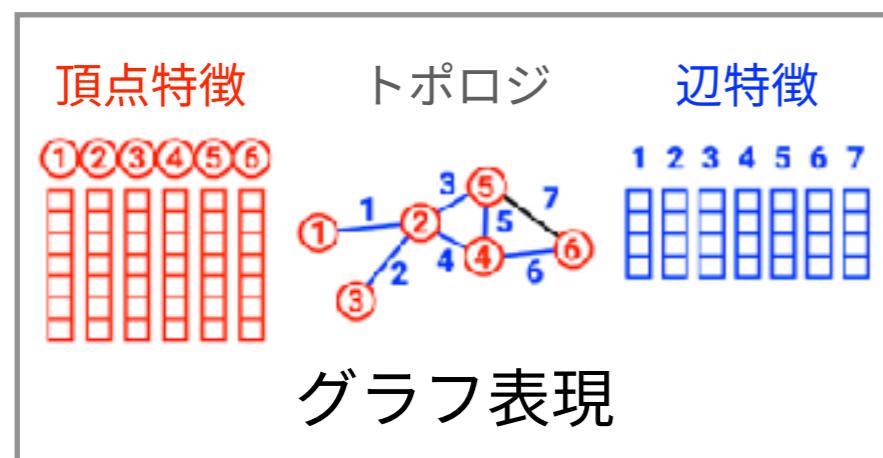
`rdkit.ML.Descriptors`

オープンソースフレームワーク

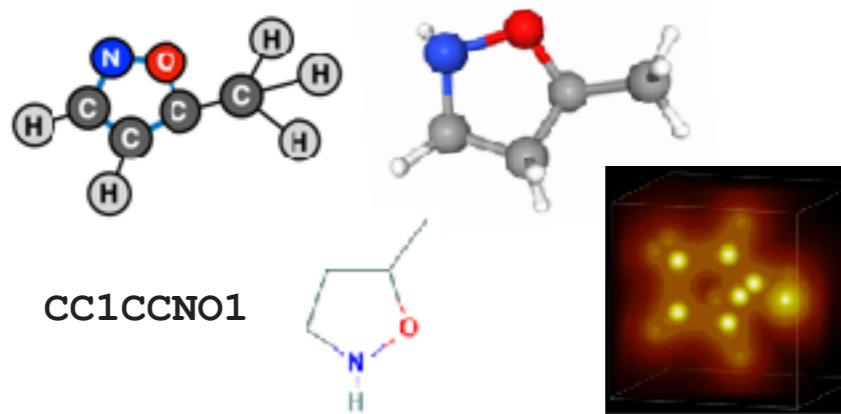
# 分子の「良い汎用的表現」の学習？



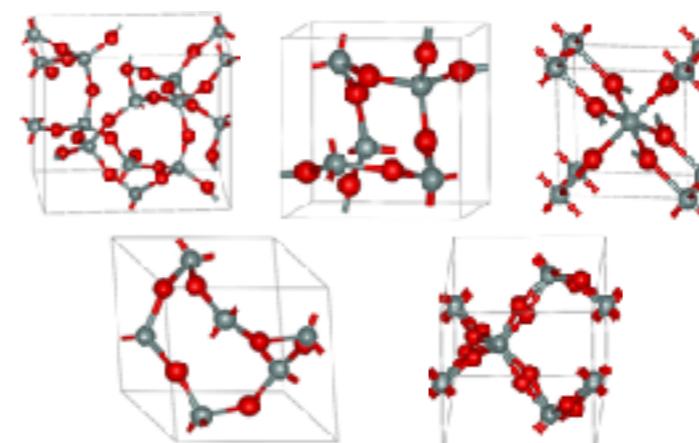
分子の環境/条件/標的/相互作用等の情報



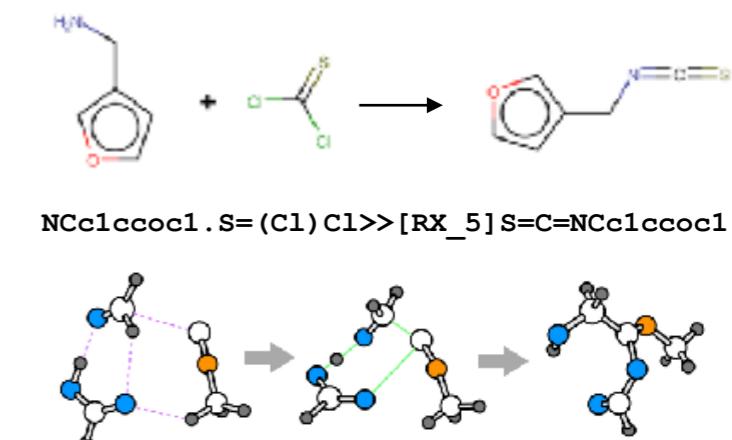
Molecules



Materials

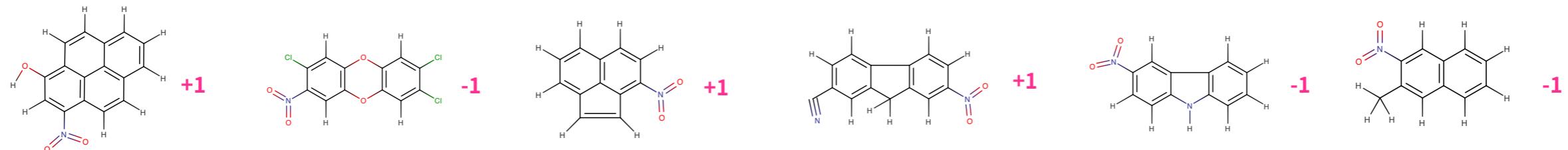


Reactions

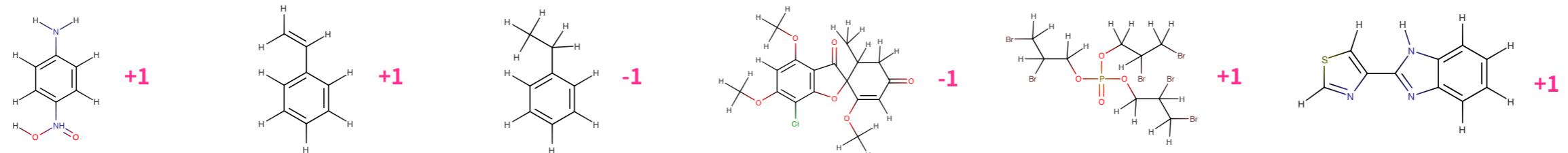


# Use Case 1: Virtual Screening (QSAR/QSPR)

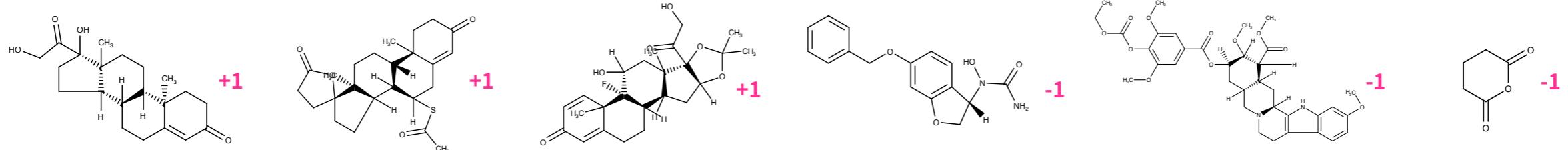
- Mutagenic potency**



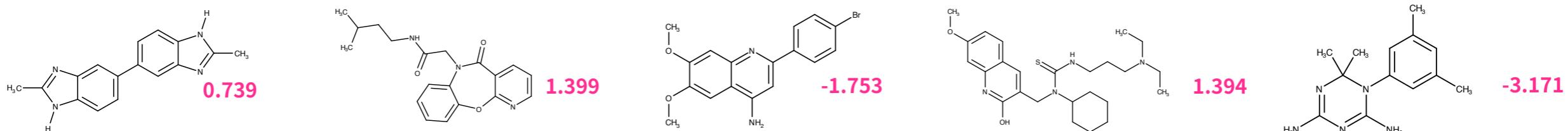
- Carcinogenic potency**



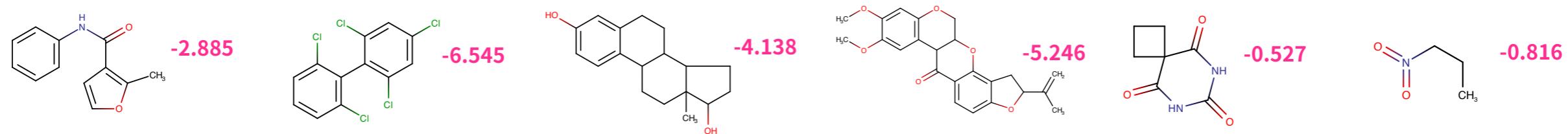
- Endocrine disruption**



- Growth inhibition**



- Aqueous solubility**



# Use Case 1: Virtual Screening (QSAR/QSPR)

<https://pubchem.ncbi.nlm.nih.gov/bioassay/1>

**BIOASSAY RECORD**

**NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Cell Lung cell line**

PubChem AID	1
Source	DTP/NCI
External ID	<a href="#">NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Cell Lung cell line</a>
BioAssay Type	Confirmatory
Tested Substances	<span>All (53,554)</span> <span>Active (3,025)</span> <span>Inactive (50,655)</span> <span>Data Table</span> 
Tested Compounds	<span>All (51,583)</span> <span>Active (2,814)</span> <span>Inactive (48,922)</span>
Version	2.1 <span>Revision History</span>
Status	Live
Dates	<span>Modify</span> 2021-07-12 <span>Deposit</span> 2004-08-15

Please note that the bioassay record (AID 1) is presented as provided to PubChem by the source(depositor). When possible, links to additional information have been provided by PubChem.

**CONTENTS**

- Title and Summary**
- 1 Description
- 2 Comment
- 3 Result Definitions
- 4 Data Table
- 5 Entrez Crosslinks
- 6 Identity
- 7 BioAssay Annotations
- 8 Information Sources

**Cite**  **Download** 

# Use Case 1: Virtual Screening (QSAR/QSPR)

input



CID 11978790

ML

output

activity: "Active"  
LogGI50: -7.8811

GI50: concentration required  
for 50% inhibition of growth

Tested Compounds

All (51,583)  Active (2,814)  Inactive (48,922)

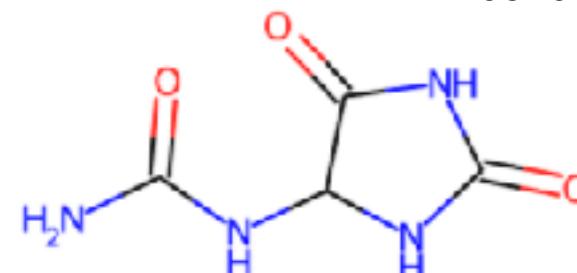
Tested Substance			Activity	Score	LogGI50_M ⓘ	LogGI50_U ⓘ	LogGI50_V ⓘ
Structure	CID	SID					
	5298	121832	<span>Active</span>	67	-8		
	363173	483713	<span>Active</span>	43	-6.5871		
	399631	530888	<span>Active</span>	51	-7.0878		
	399630	530867	<span>Active</span>	60	-7.617		

Tested Substance			Activity	Score	LogGI50_M ⓘ	LogGI50_U ⓘ	LogGI50_V ⓘ
Structure	CID	SID					
	380324	521801	<span>Inactive</span>	0	-4		
	380311	521588	<span>Inactive</span>	0	-4		
	390312	521589	<span>Inactive</span>	4	-4.214		
	135489878	521590	<span>Inactive</span>	13	-4.7552		

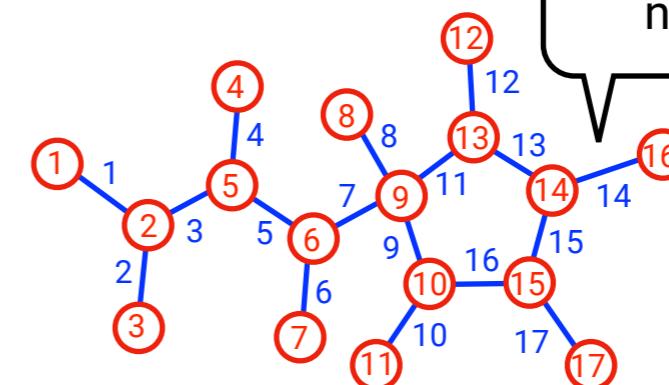
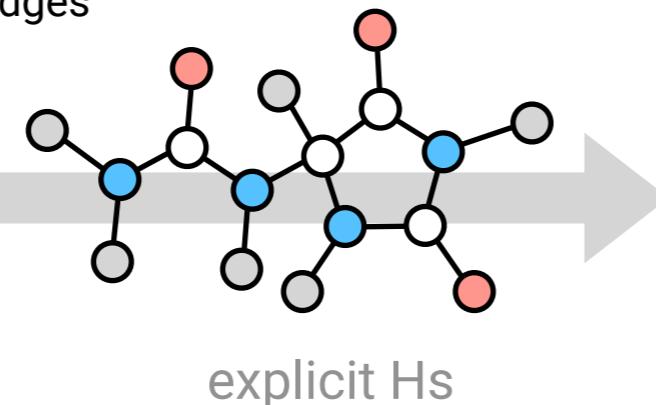
# Molecular Graphs: 分子のグラフ表現

Input representation (molecular graph)

atoms → nodes  
bonds → edges



CID 204



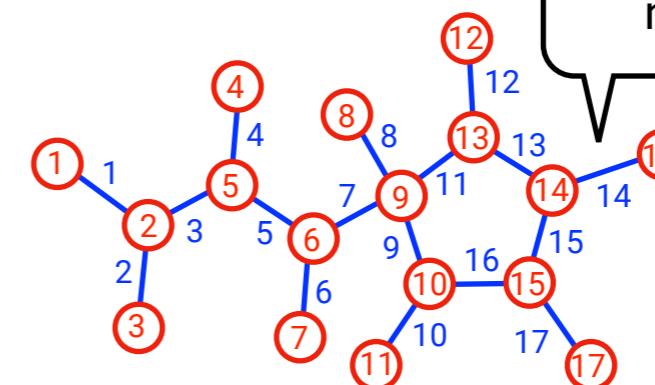
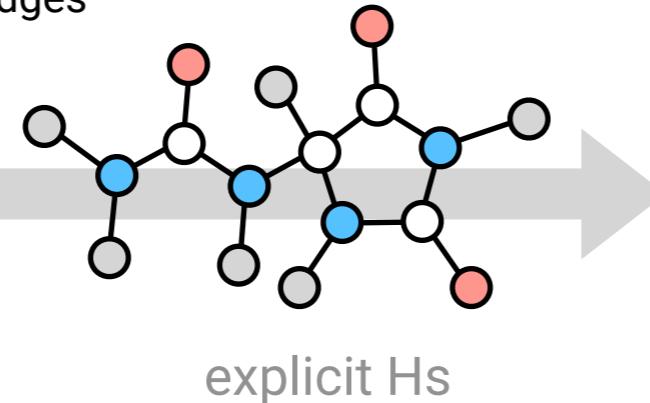
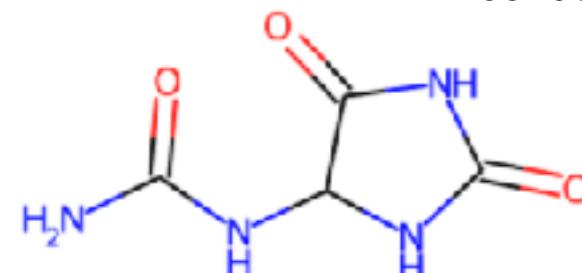
! Any permutation of this numbering should not change the results.

1. *permutation equivariance*
2. *permutation invariance*

# Molecular Graphs: 分子のグラフ表現

Input representation (molecular graph)

atoms → nodes  
bonds → edges

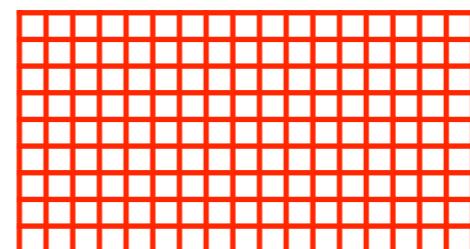


1. *permutation equivariance*
2. *permutation invariance*

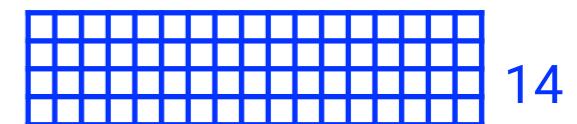
e.g. Features for ChemProp (Yang et al, 2019)

node(atom) features

17



133



edge(bond) features

17

14

133 features

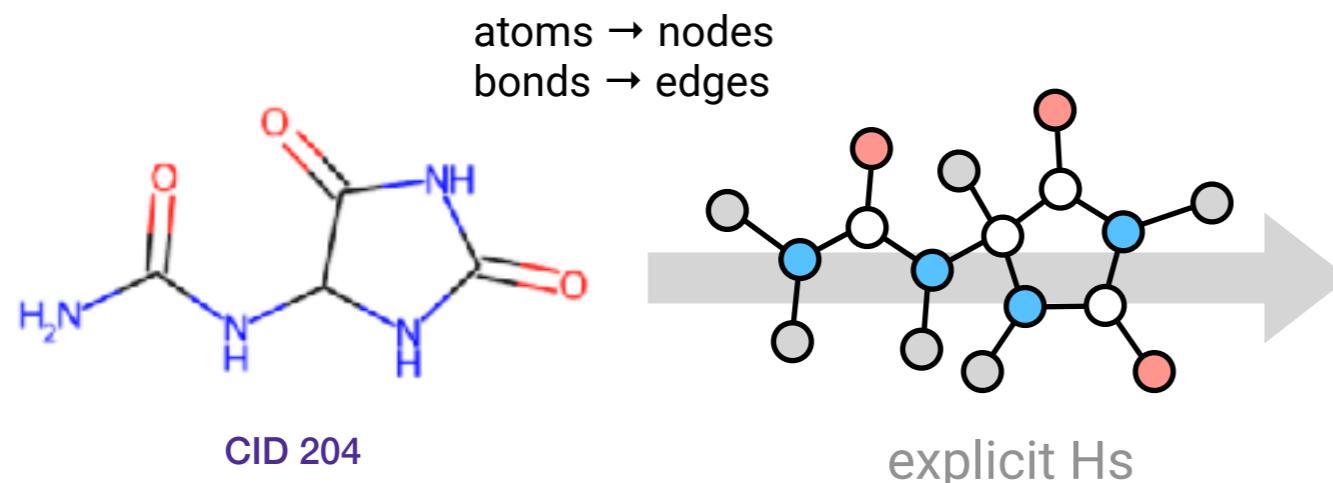
- atomic\_num (one-hot, 101)
- total\_degree (one-hot, 7)
- formal\_charge (one-hot, 6)
- chiral\_tag (one-hot, 5)
- num\_Hs (one-hot, 6)
- hybridization (one-hot, 6)
- is\_aromatic (binary, 1)
- atomic\_mass (real, 1)

14 features

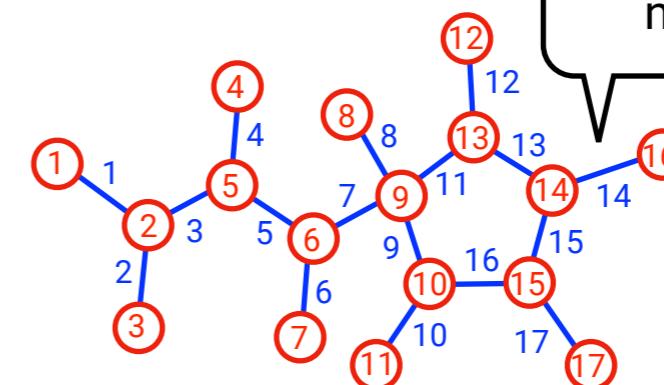
- no\_bond (binary, 1)
- is\_single (binary, 1)
- is\_double (binary, 1)
- is\_triple (binary, 1)
- is\_aromatic (binary, 1)
- is\_connjugated (binary, 1)
- is\_in\_ring (binary, 1)
- stereo (one-hot, 7)

# Molecular Graphs: 分子のグラフ表現

Input representation (molecular graph)

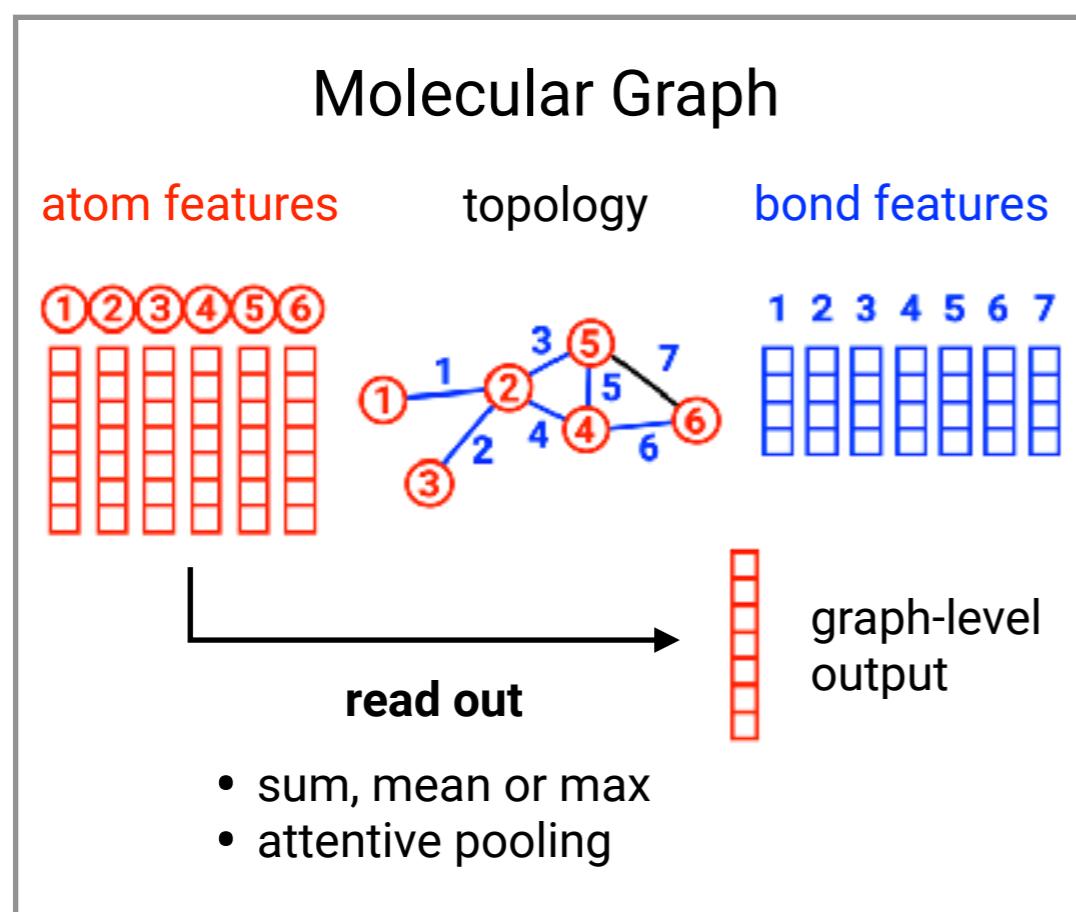


! Any permutation of this numbering should not change the results.



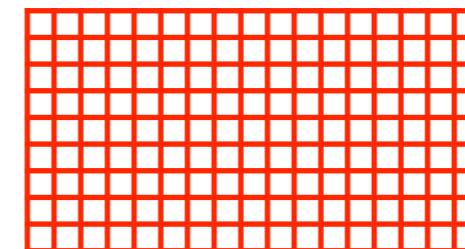
1. *permutation equivariance*
2. *permutation invariance*

e.g. Features for ChemProp (Yang et al, 2019)



**node(atom) features**

17



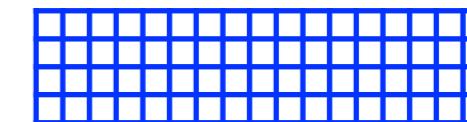
133

**133 features**

- atomic\_num (one-hot, 101)
- total\_degree (one-hot, 7)
- formal\_charge (one-hot, 6)
- chiral\_tag (one-hot, 5)
- num\_Hs (one-hot, 6)
- hybridization (one-hot, 6)
- is\_aromatic (binary, 1)
- is\_connjugated (binary, 1)
- is\_in\_ring (binary, 1)
- stereo (one-hot, 7)

**edge(bond) features**

17

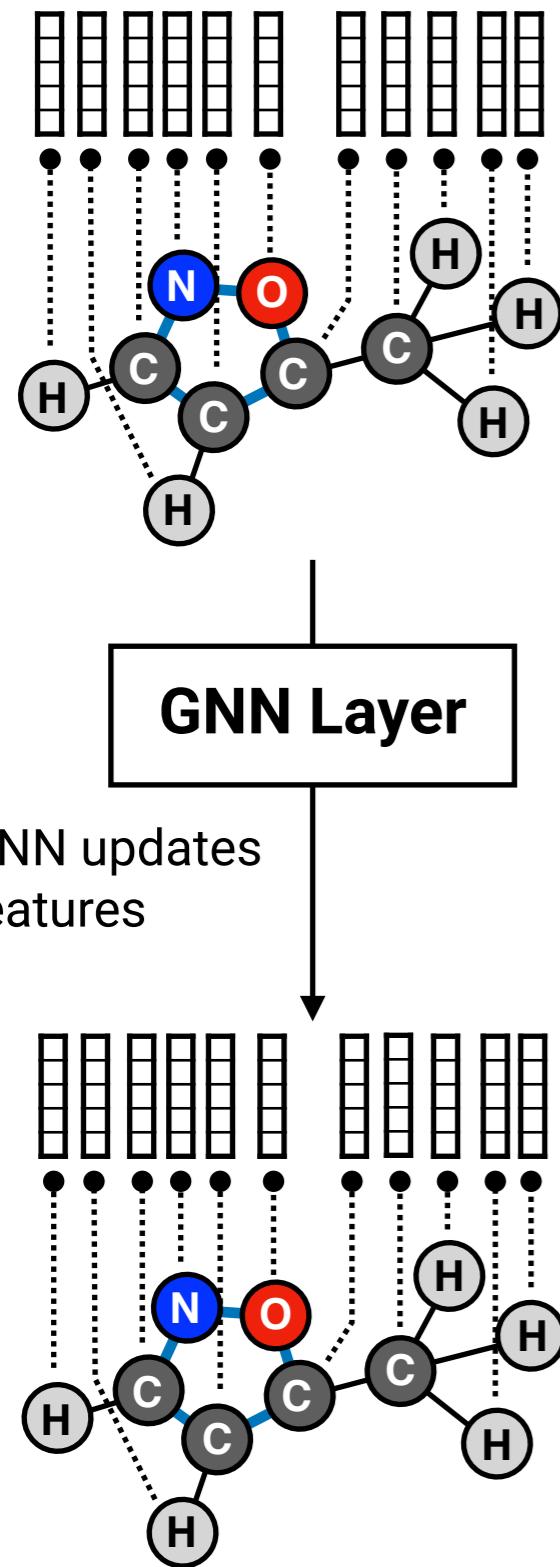


14

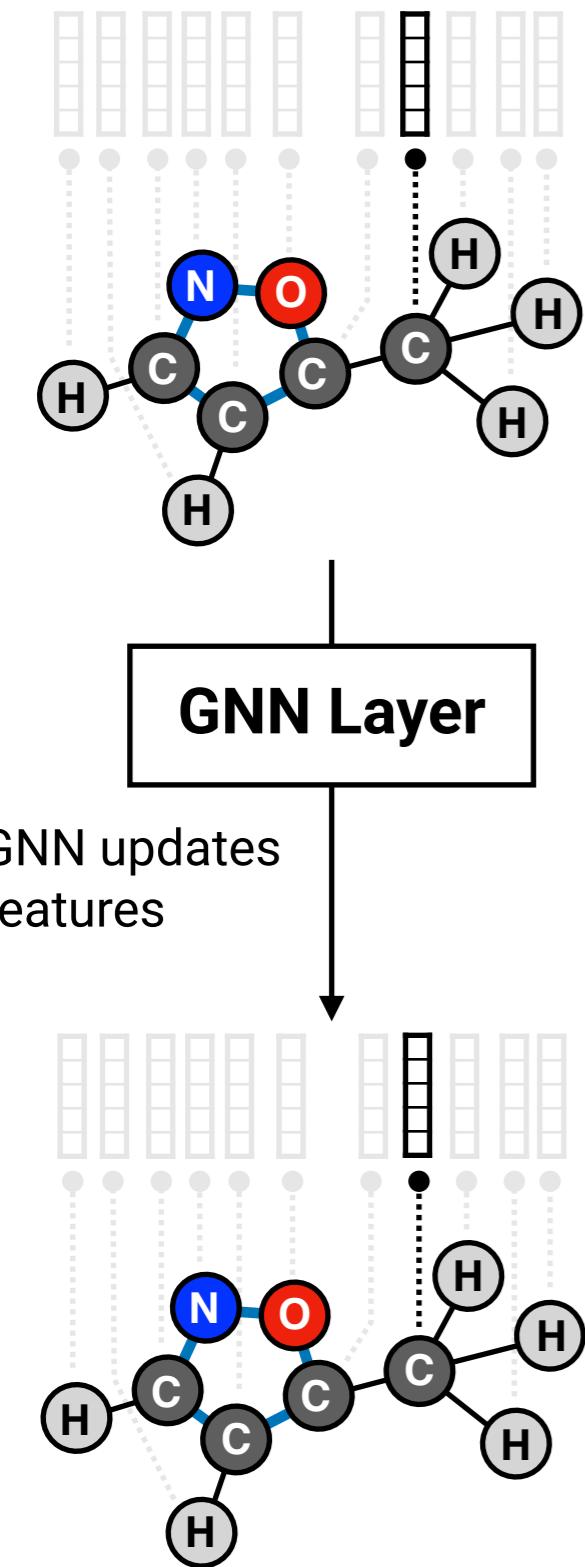
**14 features**

- no\_bond (binary, 1)
- is\_single (binary, 1)
- is\_double (binary, 1)
- is\_triple (binary, 1)
- is\_aromatic (binary, 1)
- is\_connjugated (binary, 1)
- is\_in\_ring (binary, 1)
- stereo (one-hot, 7)

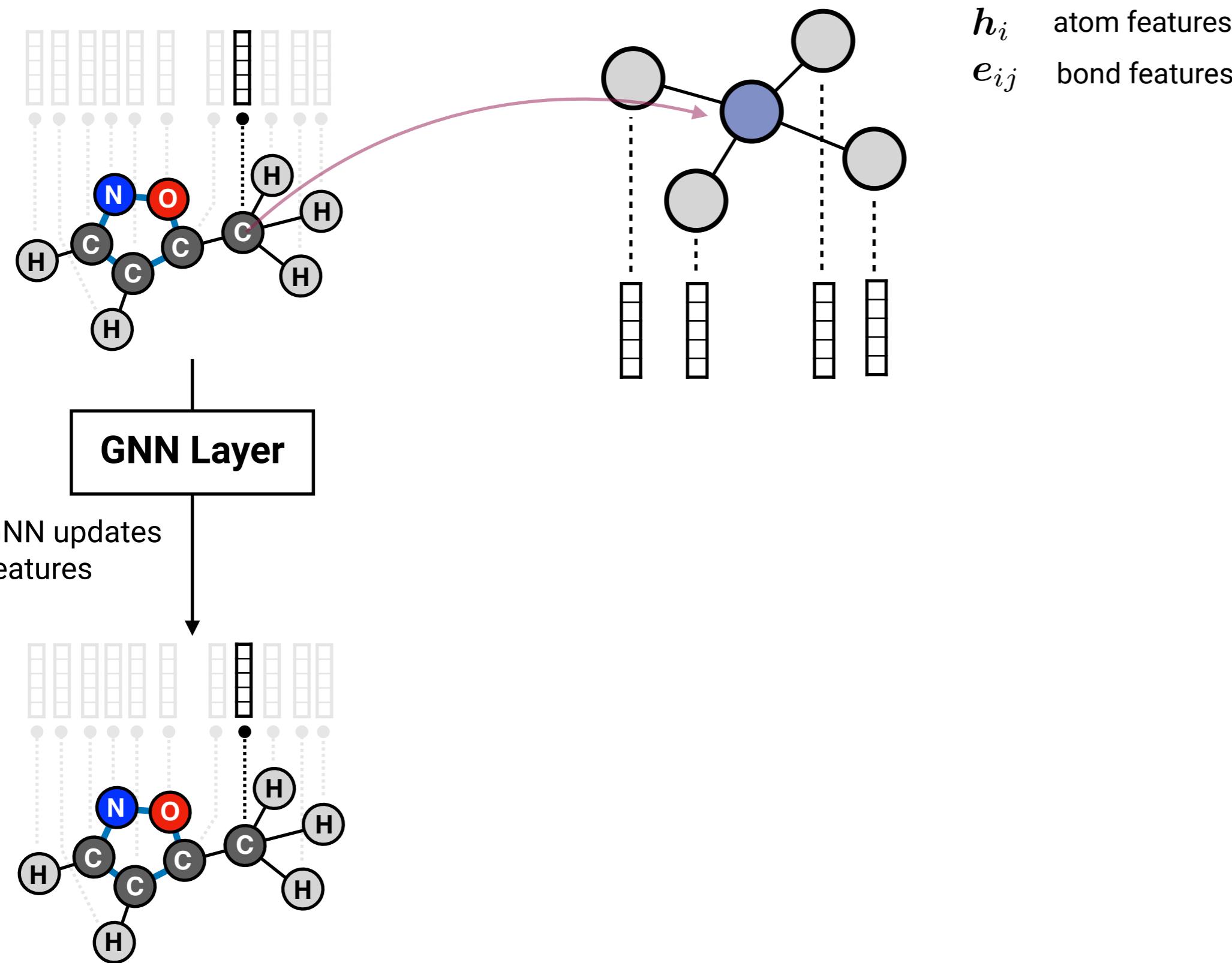
# Graph Neural Networks (GNNs)



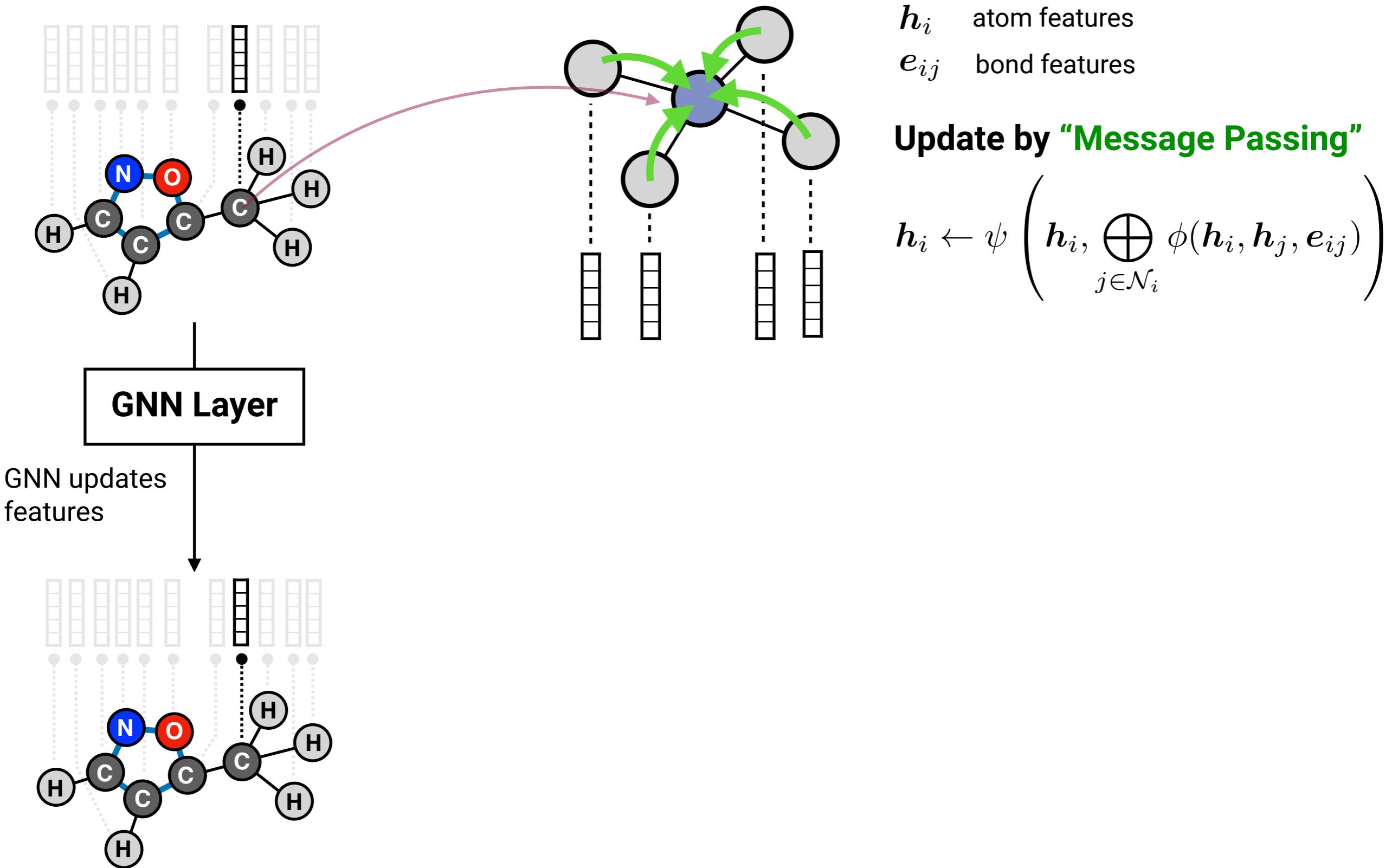
# Graph Neural Networks (GNNs)



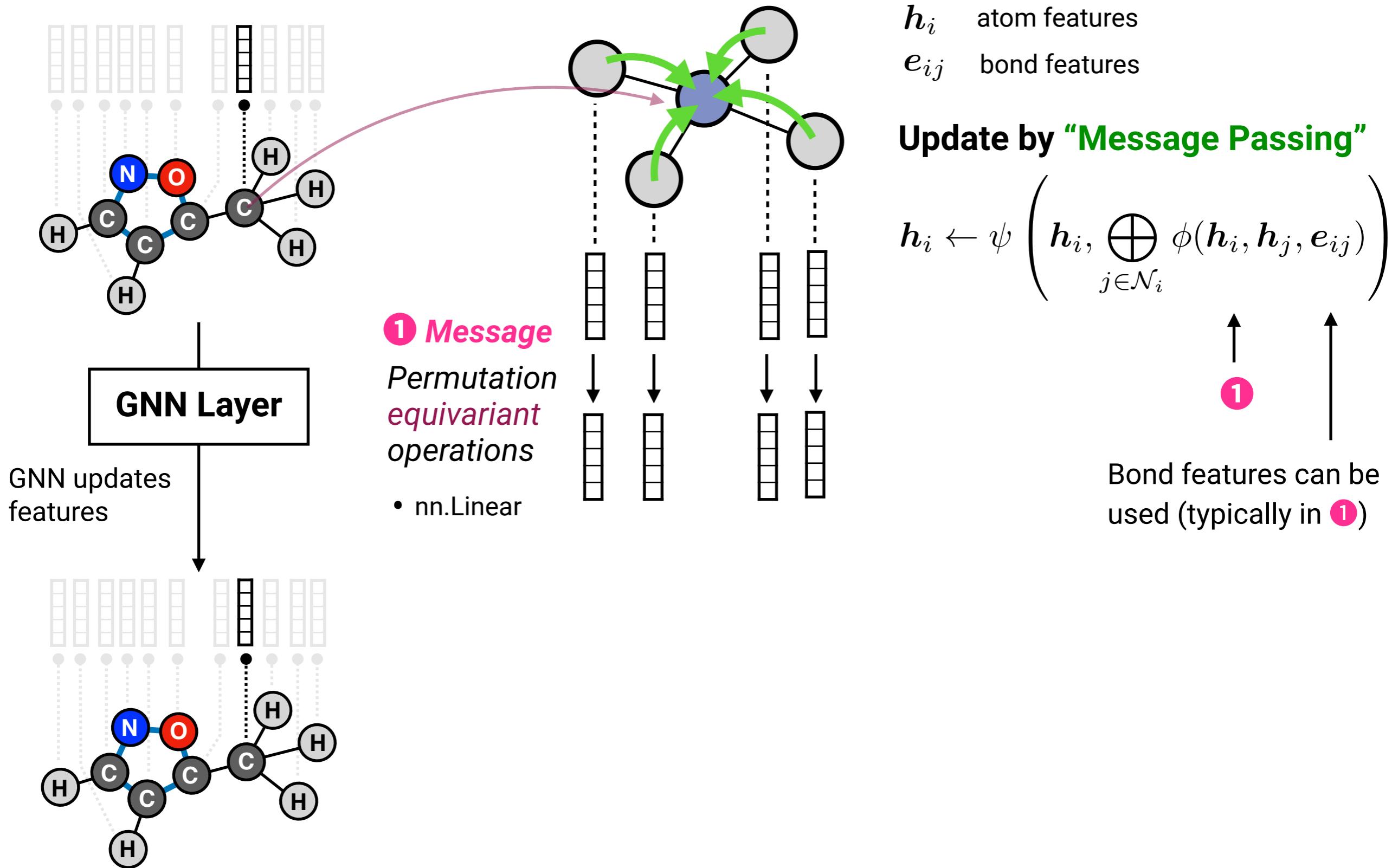
# Graph Neural Networks (GNNs)



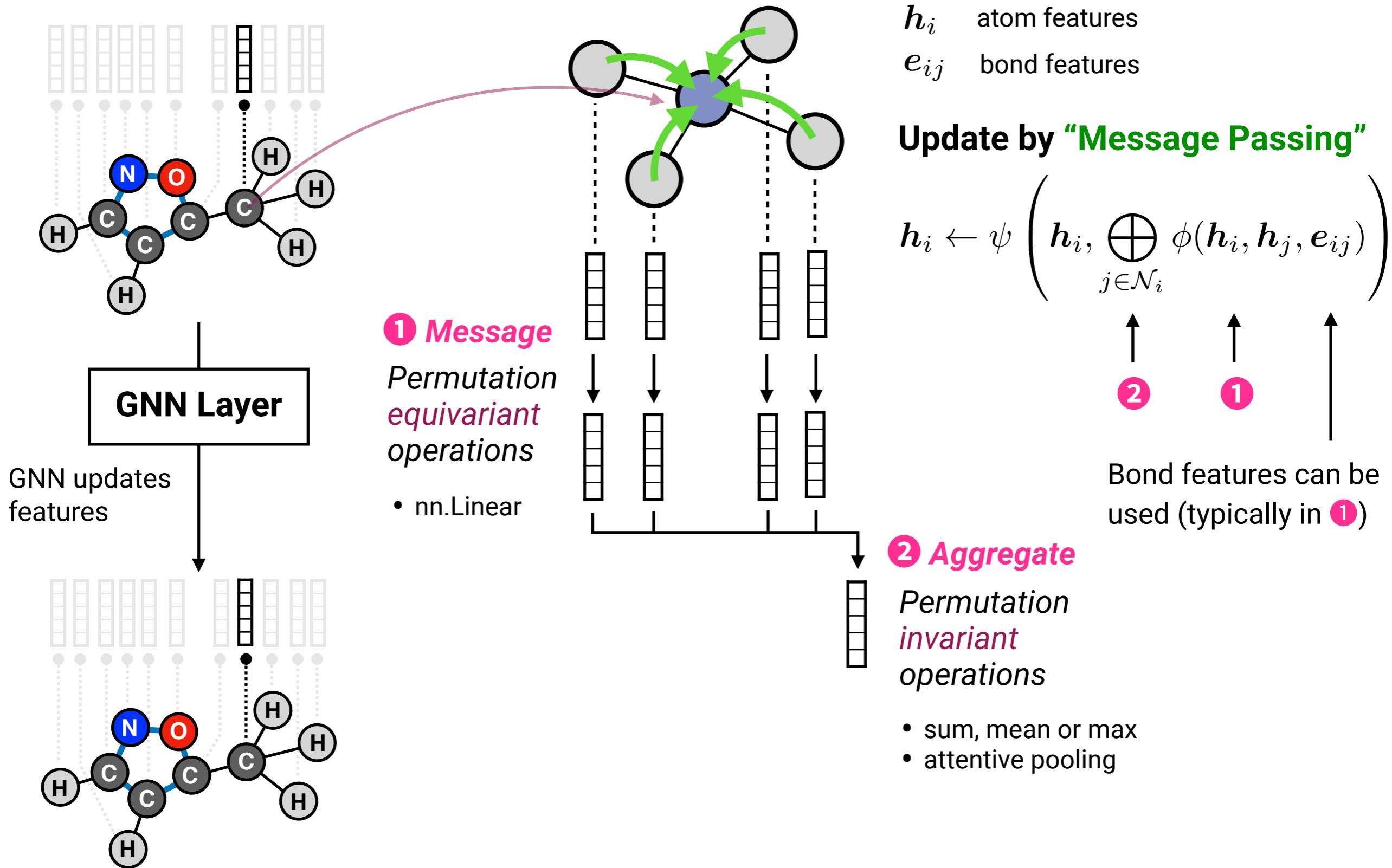
# Graph Neural Networks (GNNs)



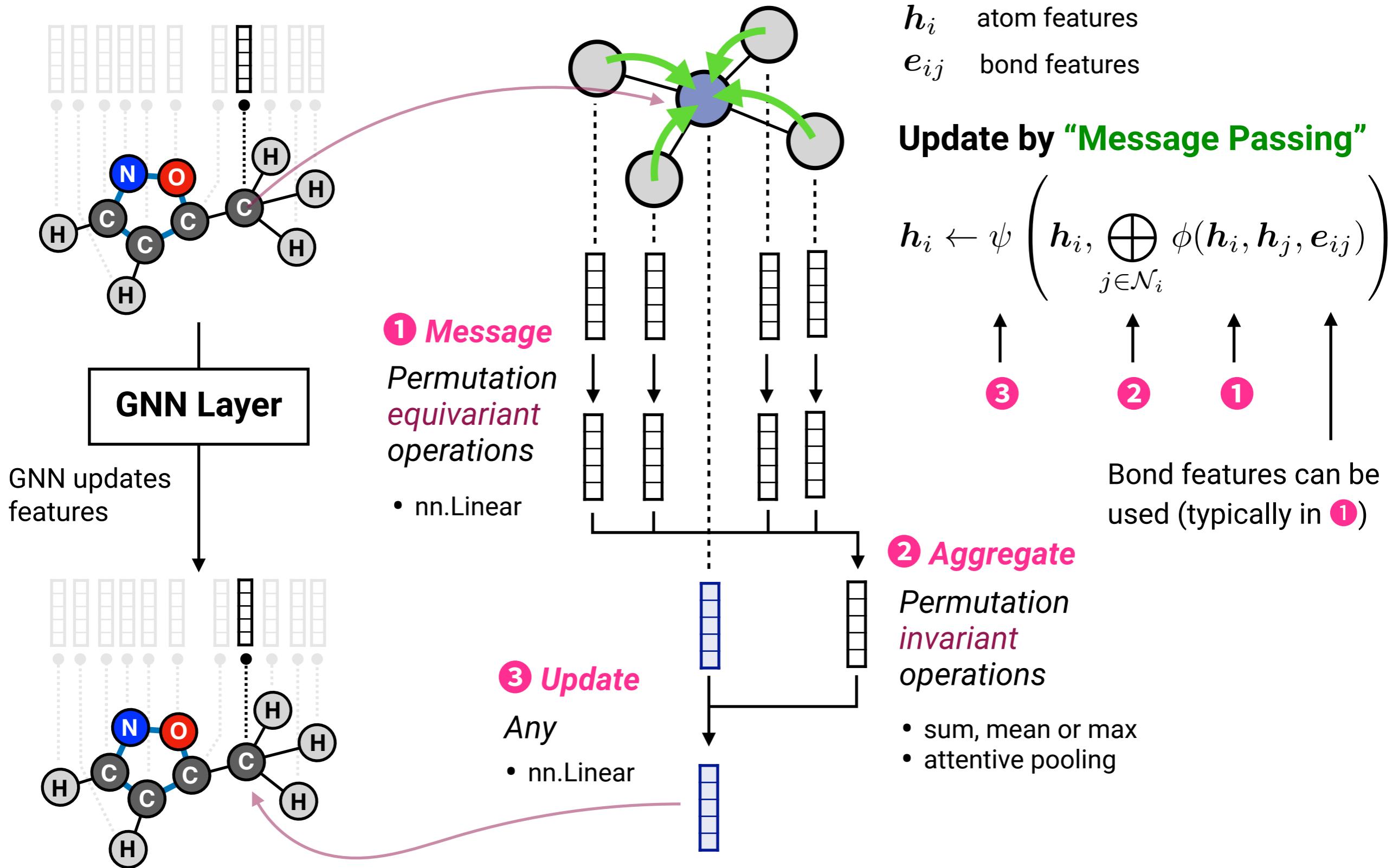
# Graph Neural Networks (GNNs)



# Graph Neural Networks (GNNs)



# Graph Neural Networks (GNNs)



# Use Case 1: Virtual Screening (QSAR/QSPR)

## Performance for **unseen (test) data:**

Active/Inactive (Classification), LogGI50 (Regression)

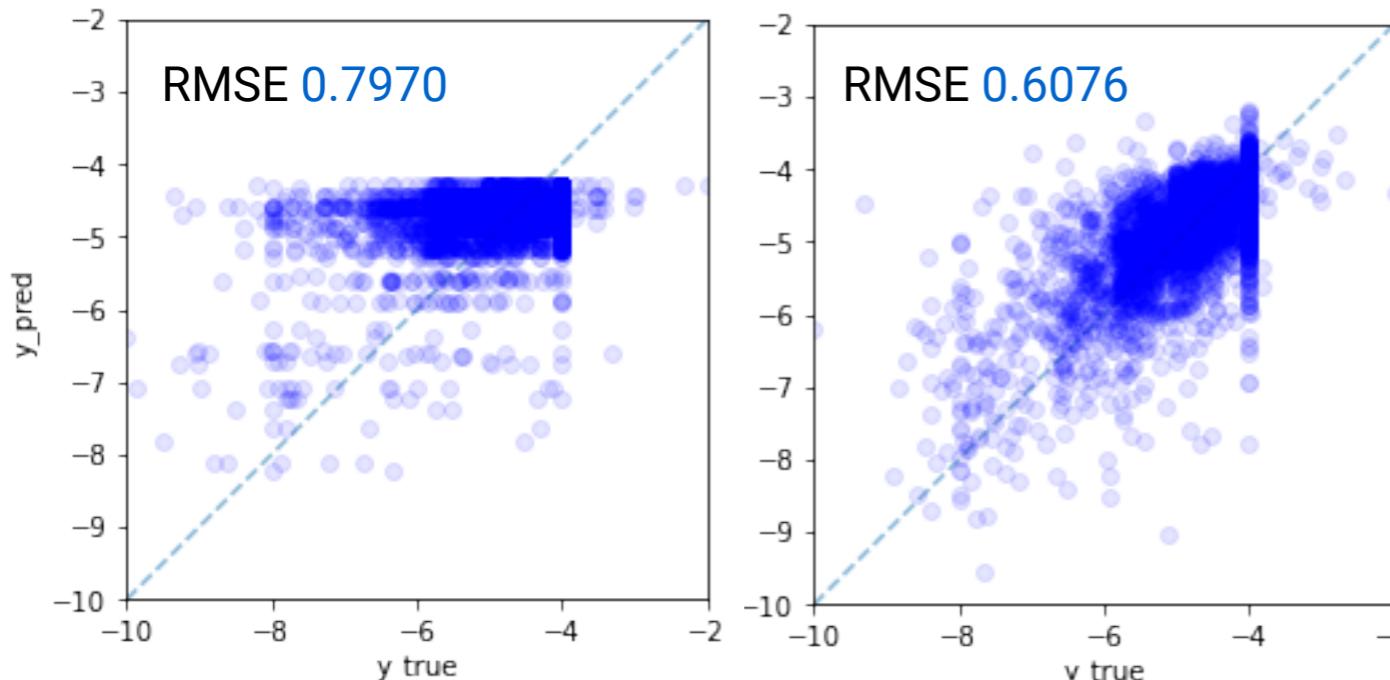
### *Standard ML*

ExtraTrees  
w/ ECFP6(1024)

- Classification accuracy

**95.079% (Active/Inactive)**

- Regression for LogGI50



*Disclaimer:* This is just for a toy demo. This should be taken as classification for ACTIVITY\_OUTCOME (Active or Inactive)

### *GNN*

ChemProp  
(Directed MPNN)

- Classification accuracy

**95.604% (Active/Inactive)**

- Regression for LogGI50

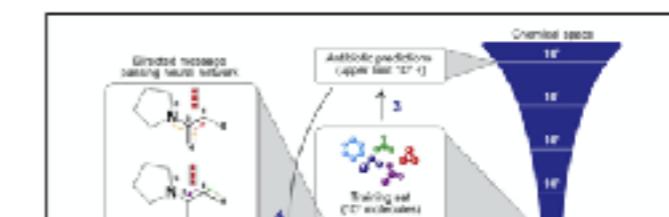
## **ChemProp (Yang et al, 2019)**

from MIT MLPDS (Machine Learning for Pharmaceutical Discovery and Synthesis) Consortium

### **Cell**

#### **A Deep Learning Approach to Antibiotic Discovery**

##### **Graphical Abstract**



##### **Authors**

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, ..., Tommi S. Jaakkola, Regina Barzilay, James J. Collins

##### **Correspondence**

regina@csail.mit.edu (R.B.), jmc@mit.edu (J.J.C.)

Stokes et al, *Cell* (2020) <https://doi.org/10.1016/j.cell.2020.01.021>

### **nature**

NEWS | 20 February 2020

#### **Powerful antibiotics discovered using AI**

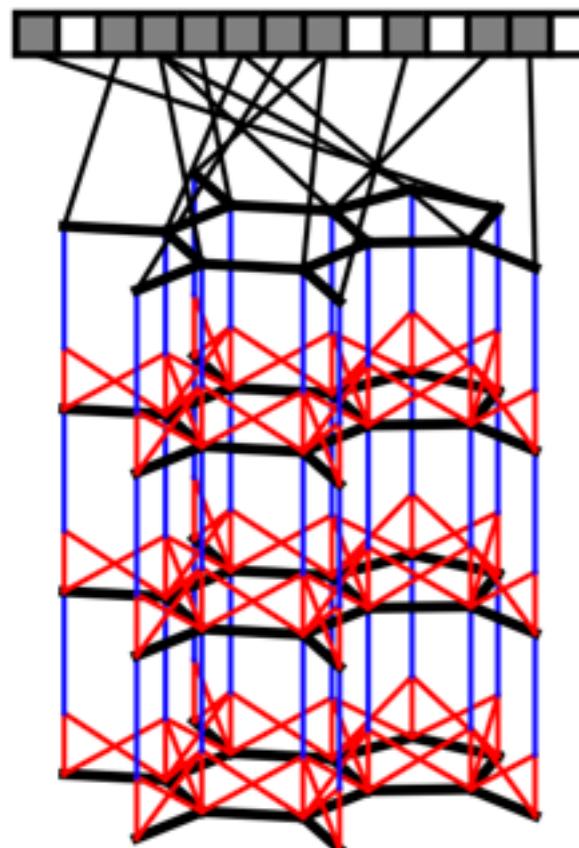
Machine learning spots molecules that work even against 'untreatable' strains of bacteria.

Jo Merchant

Merchant, *Nature* (2020) <https://doi.org/10.1038/d41586-020-00018-3>

# ECFPとNeural Graph Fingerprint

- Neural Graph Fingerprint: 最初期に提案されたGNNの一つ
- Graph Convolutionを用いたGNNの一種とみなせる
- ECFP(Circular Fingerprint)のFingerprint計算をパラメタを持つ微分可能な演算で書き直すことで得られる学習可能なFingerprintという位置づけ




---

**Algorithm 1 Circular fingerprints**


---

```

1: Input: molecule, radius  $R$ , fingerprint length  $S$ 
2: Initialize: fingerprint vector  $\mathbf{f} \leftarrow \mathbf{0}_S$ 
3: for each atom  $a$  in molecule
4:    $\mathbf{r}_a \leftarrow g(a)$             $\triangleright$  lookup atom features
5:   for  $L = 1$  to  $R$             $\triangleright$  for each layer
6:     for each atom  $a$  in molecule
7:        $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$ 
8:        $\mathbf{v} \leftarrow [\mathbf{r}_a, \mathbf{r}_1, \dots, \mathbf{r}_N]$      $\triangleright$  concatenate
9:        $\mathbf{r}_a \leftarrow \text{hash}(\mathbf{v})$             $\triangleright$  hash function
10:       $i \leftarrow \text{mod}(r_a, S)$          $\triangleright$  convert to index
11:       $\mathbf{f}_i \leftarrow 1$                    $\triangleright$  Write 1 at index
12: Return: binary vector  $\mathbf{f}$ 
```

---



---

**Algorithm 2 Neural graph fingerprints**


---

```

1: Input: molecule, radius  $R$ , hidden weights  $H_1^1 \dots H_R^5$ , output weights  $W_1 \dots W_R$ 
2: Initialize: fingerprint vector  $\mathbf{f} \leftarrow \mathbf{0}_S$ 
3: for each atom  $a$  in molecule
4:    $\mathbf{r}_a \leftarrow g(a)$             $\triangleright$  lookup atom features
5:   for  $L = 1$  to  $R$             $\triangleright$  for each layer
6:     for each atom  $a$  in molecule
7:        $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$ 
8:        $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^N \mathbf{r}_i$            $\triangleright$  sum
9:        $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$             $\triangleright$  smooth function
10:       $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$          $\triangleright$  sparsify
11:       $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$                    $\triangleright$  add to fingerprint
12: Return: real-valued vector  $\mathbf{f}$ 
```

---

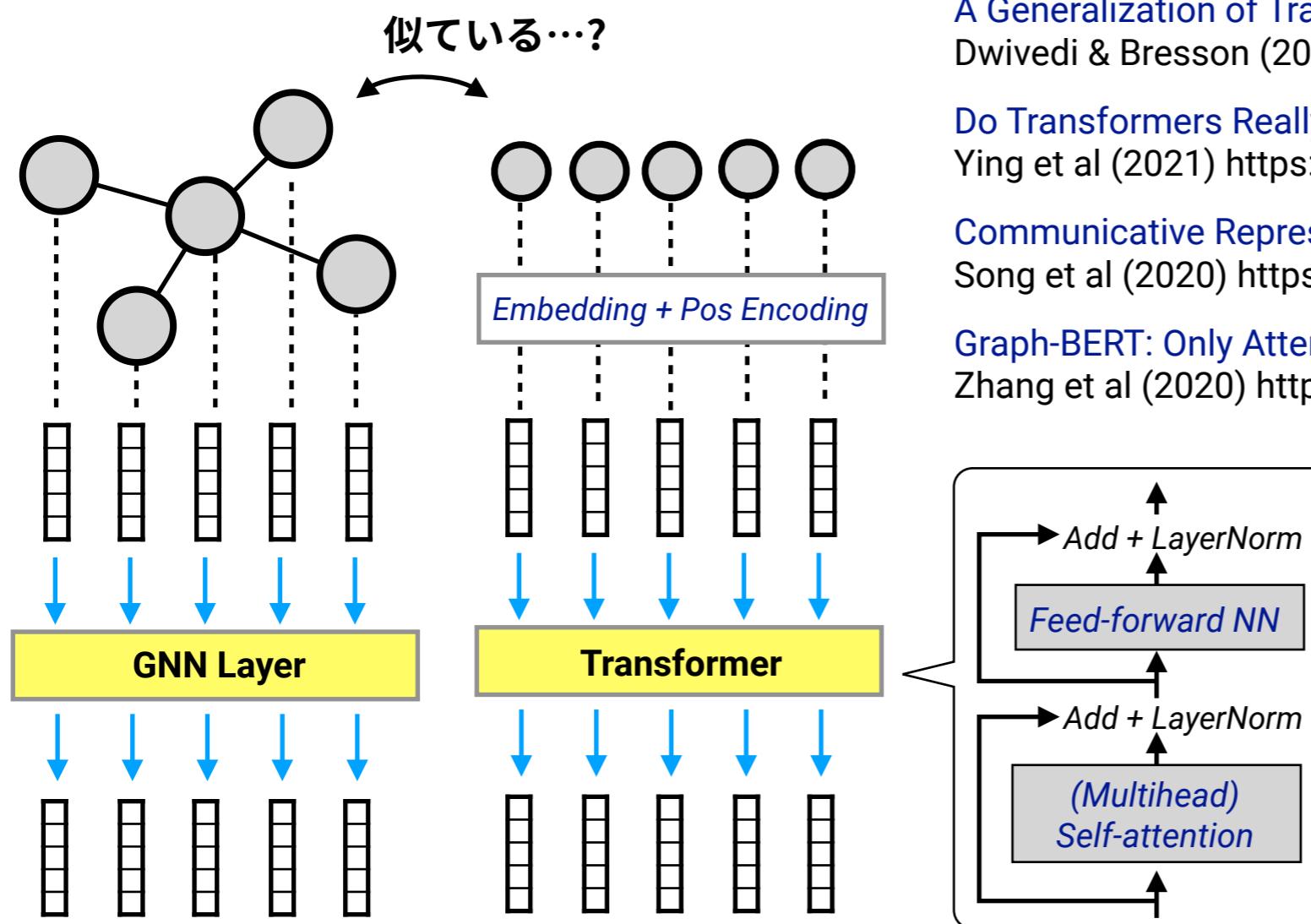
Figure 2: Pseudocode of circular fingerprints (*left*) and neural graph fingerprints (*right*). Differences are highlighted in blue. Every non-differentiable operation is replaced with a differentiable analog.

# GATとTransformer型GNN

- 各頂点の特徴ベクトルを更新する際にAttentionを入れたい
- Transformerはトポロジ制約のないGraph Attention Network (GAT)変種とみなせる

Veličković, Cucurull, Casanova, Romero, Liò, Bengio, [Graph Attention Networks](https://arxiv.org/abs/1710.10903) (ICLR 2018) <https://arxiv.org/abs/1710.10903>  
Joshi, [Transformers are Graph Neural Networks.](https://graphdeeplearning.github.io/post/transformers-are-gnns/) (2020) <https://graphdeeplearning.github.io/post/transformers-are-gnns/>

- 逆にもちろんTransformer型のSelf-AttentionをGNNにもちこむこともできる



A Generalization of Transformer Networks to Graphs  
Dwivedi & Bresson (2020) <https://arxiv.org/abs/2012.09699>

Do Transformers Really Perform Bad for Graph Representation?  
Ying et al (2021) <https://arxiv.org/abs/2106.05234>

Communicative Representation Learning on Attributed Molecular Graphs  
Song et al (2020) <https://www.ijcai.org/proceedings/2020/0392.pdf>

Graph-BERT: Only Attention is Needed for Learning Graph Representations  
Zhang et al (2020) <https://arxiv.org/abs/2001.05140>

Ying et al (2021) のGraphomerは  
KDDCup 2021のOpen Graph Benchmark  
Large-Scale Challenge(後述)のGraph-level  
タスクの優勝モデルで使われた

大規模データならグラフでも  
Transformerは有効...!?

# 分子表現の事前学習と転移学習

- Transformerへの関心は(Self-Supervisedな)大規模事前学習と転移への期待の現れ
- 分子タスクも現実の個別状況では小サンプルであることがほとんど
- もし汎用の分子表現を大規模事前学習により獲得しFew-shot/Zero-shot転移ができるのなら波及効果は計り知れない (cf. CVのImageNet-pretrained CNN, NLPのBERT等)

## Self-Supervised Graph Transformer on Large-Scale Molecular Data

Rong, Bian, Xu, Xie, Wei, Huang, Huang (NeurIPS 2020)

<https://arxiv.org/abs/2007.02835>

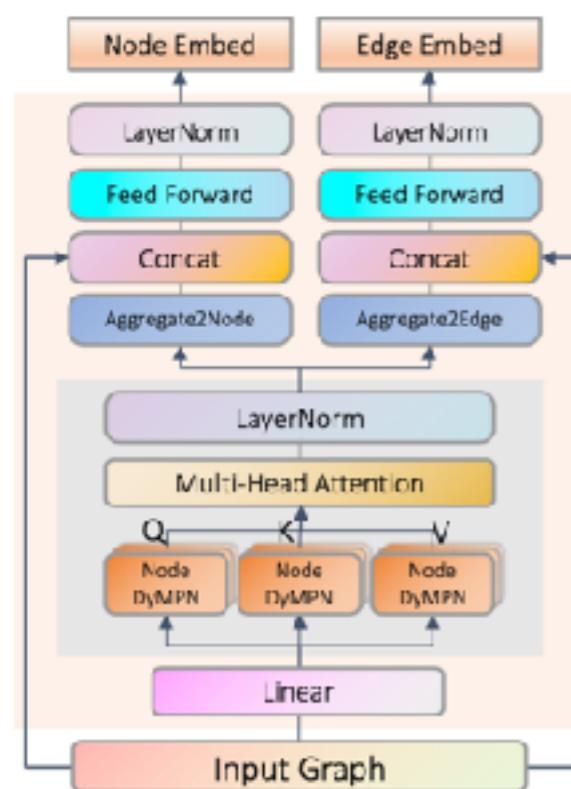


Figure 1: Overview of GTransformer.

## Strategies for Pre-training Graph Neural Networks

Hu, Liu, Gomes, Zitnik, Liang, Pande, Leskovec (ICLR 2020)

<https://arxiv.org/abs/1905.12265>

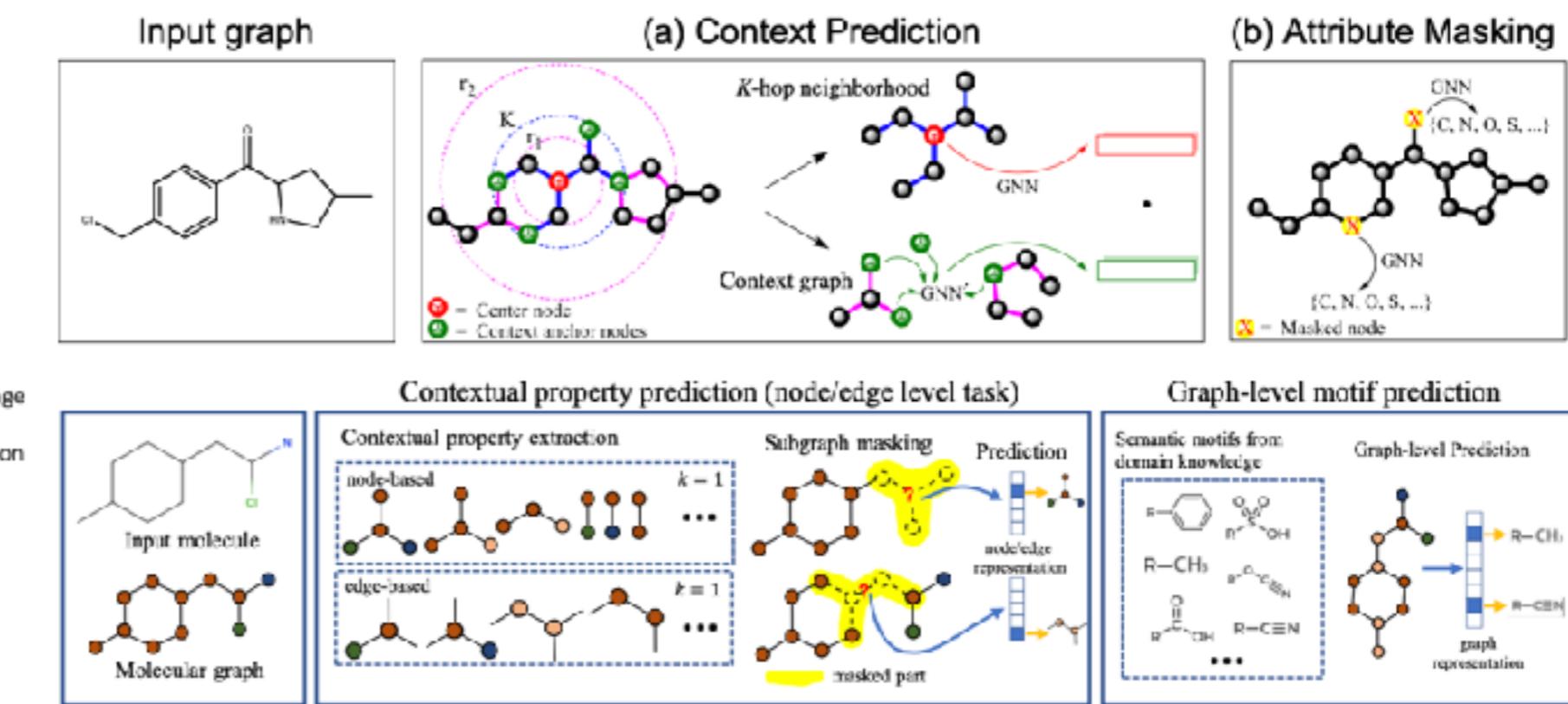


Figure 2: Overview of the designed self-supervised tasks of GROVER.

# 分子表現の生成

- もうひとつの分子の表現学習への期待は**分子グラフや分子構造の生成**
- 分子生成の場合は特にDecoderが非自明で構造的な処理を実現する必要がある
- 構成性/モジュール性や化学的ルールも考慮しないと意味のない出力になり得る
- 文字列表現(SMILES記法)からの生成は直接的なのでグラフ表現の優位性も要検証

## Deep Graph Generators: A Survey

FAEZEH FAEZ<sup>1</sup>, YASSAMAN OMMI<sup>2</sup>, MAHDIEH SOLEYMANI BAGHSHAH<sup>1</sup>, AND HAMID R. RABIEE<sup>1</sup>, (Senior Member, IEEE)

<sup>1</sup>Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

<sup>2</sup>Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran

Corresponding authors: Hamid R. Rabiee and Mahdiedh Soleymani Baghshah (e-mails: rabiee@sharif.edu , soleymani@sharif.edu).

Category	Key Characteristic	Publications
Autoregressive DGGs	Adopting a sequential generation strategy, either node-by-node or edge-by-edge	[1]–[26]
Autoencoder-Based DGGs	Making the generation process dependent on latent space variables	[14]–[19], [27]–[39]
RL-Based DGGs	Utilizing reinforcement learning algorithms to induce desired properties in the generated graphs	[3], [20]–[26], [40]
Adversarial DGGs	Employing generative adversarial networks (GANs) [41] to generate graph structures	[20], [22], [38]–[40], [42]–[47]
Flow-based DGGs	Learning a mapping from the complicated graph distribution into a distribution mostly modeled as a Gaussian for calculating the exact data likelihood	[12], [13], [37], [48]

# Use Case 2: Quantum chemistry

[https://qcarchive.molssi.org/apps/ml\\_datasets/](https://qcarchive.molssi.org/apps/ml_datasets/)


Machine Learning Datasets Repository

Search:  ×

[Add your Dataset](#) [License](#)

Name	Quality	Data Points	Elements	Sampling	Download
+ ANI-1	DFT	22,057,374	<span style="border: 1px solid black; padding: 2px 5px;">C</span> <span style="border: 1px solid black; padding: 2px 5px;">H</span> <span style="border: 1px solid black; padding: 2px 5px;">N</span> <span style="border: 1px solid black; padding: 2px 5px;">O</span>	NMS	<span style="color: blue; border: 1px solid black; padding: 2px 5px;">↓ HDFS</span> <span style="color: red; border: 1px solid black; padding: 2px 5px;">↓ TEXT</span>
+ ANI-1x	DFT	4,956,005	<span style="border: 1px solid black; padding: 2px 5px;">C</span> <span style="border: 1px solid black; padding: 2px 5px;">H</span> <span style="border: 1px solid black; padding: 2px 5px;">N</span> <span style="border: 1px solid black; padding: 2px 5px;">O</span>	MD,NMS,DS,TS	<span style="color: blue; border: 1px solid black; padding: 2px 5px;">↓ HDFS</span>
- QM9	DFT	133,885	<span style="border: 1px solid black; padding: 2px 5px;">C</span> <span style="border: 1px solid black; padding: 2px 5px;">H</span> <span style="border: 1px solid black; padding: 2px 5px;">F</span> <span style="border: 1px solid black; padding: 2px 5px;">N</span> <span style="border: 1px solid black; padding: 2px 5px;">O</span>	Minima	<span style="color: blue; border: 1px solid black; padding: 2px 5px;">↓ HDFS</span> <span style="color: red; border: 1px solid black; padding: 2px 5px;">↓ TEXT</span>

Description

Small organic molecules with up to 9 heavy atoms sampled from GDB-17, optimized at the B3LYP/6-31G(2df,p) level of theory. Ground state, orbital, and thermodynamic properties are available (at the B3LYP/6-31G(2df,p) level). All molecules are neutral singlets. This dataset was sourced from [quantum-machine.org](#) and [qmml.org](#).

Elements: C H F N O

Labels

energy homo lumo polarizability dipole frequency zpve  
enthalpy free energy heat capacity rotational constant

Tags

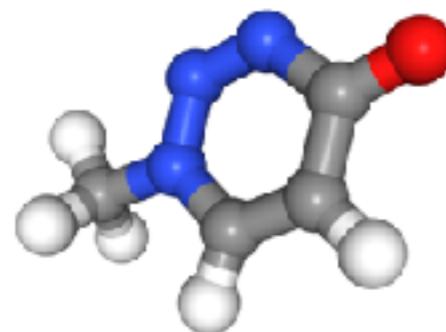
organic thermodynamics GDB

Citations

- Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13.

# Use Case 2: Quantum chemistry

input



gdb\_21014

	x	y	z
O	0.314096	-0.129589	-0.389150
C	0.111219	2.102676	-0.051749
C	2.331344	3.941075	0.212303
O	4.667017	2.677399	0.437948
C	6.152491	3.062553	-1.780599
C	4.732264	5.009654	-3.282819
C	2.562527	5.549427	-2.143825
H	-1.771427	3.048695	0.071772
H	1.977918	5.086871	1.919865
H	8.050245	3.696867	-1.222422
H	6.372399	1.276980	-2.825015
H	5.428656	5.805758	-5.033531
H	1.118529	6.857080	-2.763050

output

	property	value
0	dipole_moment	7.214000
1	isotropic_polarizability	65.360001
2	homo	-6.280388
3	lumo	-1.649010
4	gap	4.631378
5	electronic_spatial_extent	884.587524
6	zpve	2.610307
7	energy_U0	-10742.250000
8	energy_U	-10742.060547
9	enthalpy_H	-10742.035156
10	free_energy_G	-10743.111328
11	heat_capacity	24.756001
12	U_0_atom	-56.213203
13	U_atomization	-56.525291
14	H_atomization	-56.833679
15	G_atomization	-52.407772
16	rotational_a	5.712810
17	rotational_b	1.644960
18	rotational_c	1.287640

~ 1000 sec

QM計算

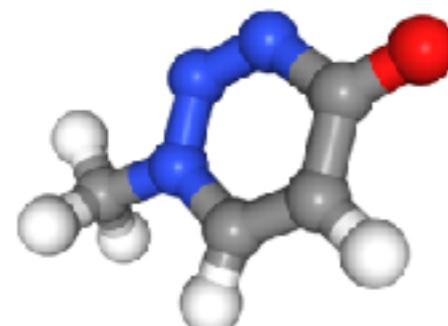
例) 一電子版のSchrödinger方程式  
(Kohn-Sham方程式)の求解

Density Functional Theory (DFT)  
B3LYP/6-31G(2df, p)

$$\hat{H}\Psi = E\Psi$$

# Use Case 2: Quantum chemistry

input



gdb\_21014

	x	y	z
O	0.314096	-0.129589	-0.389150
C	0.111219	2.102676	-0.051749
C	2.331344	3.941075	0.212303
O	4.667017	2.677399	0.437948
C	6.152491	3.062553	-1.780599
C	4.732264	5.009654	-3.282819
C	2.562527	5.549427	-2.143825
H	-1.771427	3.048695	0.071772
H	1.977918	5.086871	1.919865
H	8.050245	3.696867	-1.222422
H	6.372399	1.276980	-2.825015
H	5.428656	5.805758	-5.033531
H	1.118529	6.857080	-2.763050

100,000 times faster!

~ 0.01 sec

ML

||

~ 1000 sec

QM計算

例) 一電子版のSchrödinger方程式  
(Kohn-Sham方程式)の求解

Density Functional Theory (DFT)  
B3LYP/6-31G(2df, p)

$$\hat{H}\Psi = E\Psi$$

output

	property	value
0	dipole_moment	7.214000
1	isotropic_polarizability	65.360001
2	homo	-6.280388
3	lumo	-1.649010
4	gap	4.631378
5	electronic_spatial_extent	884.587524
6	zpve	2.610307
7	energy_U0	-10742.250000
8	energy_U	-10742.060547
9	enthalpy_H	-10742.035156
10	free_energy_G	-10743.111328
11	heat_capacity	24.756001
12	U_0_atom	-56.213203
13	U_atomization	-56.525291
14	H_atomization	-56.833679
15	G_atomization	-52.407772
16	rotational_a	5.712810
17	rotational_b	1.644960
18	rotational_c	1.287640

# Use Case 2: Quantum chemistry

- Googleが色々なGNNのバリエーションを「MPNN(Message Passing NN)」として統一的に見直した際にターゲットにされたのがこの量子化学計算近似タスク

ICML 2017 <https://arxiv.org/abs/1704.01212>

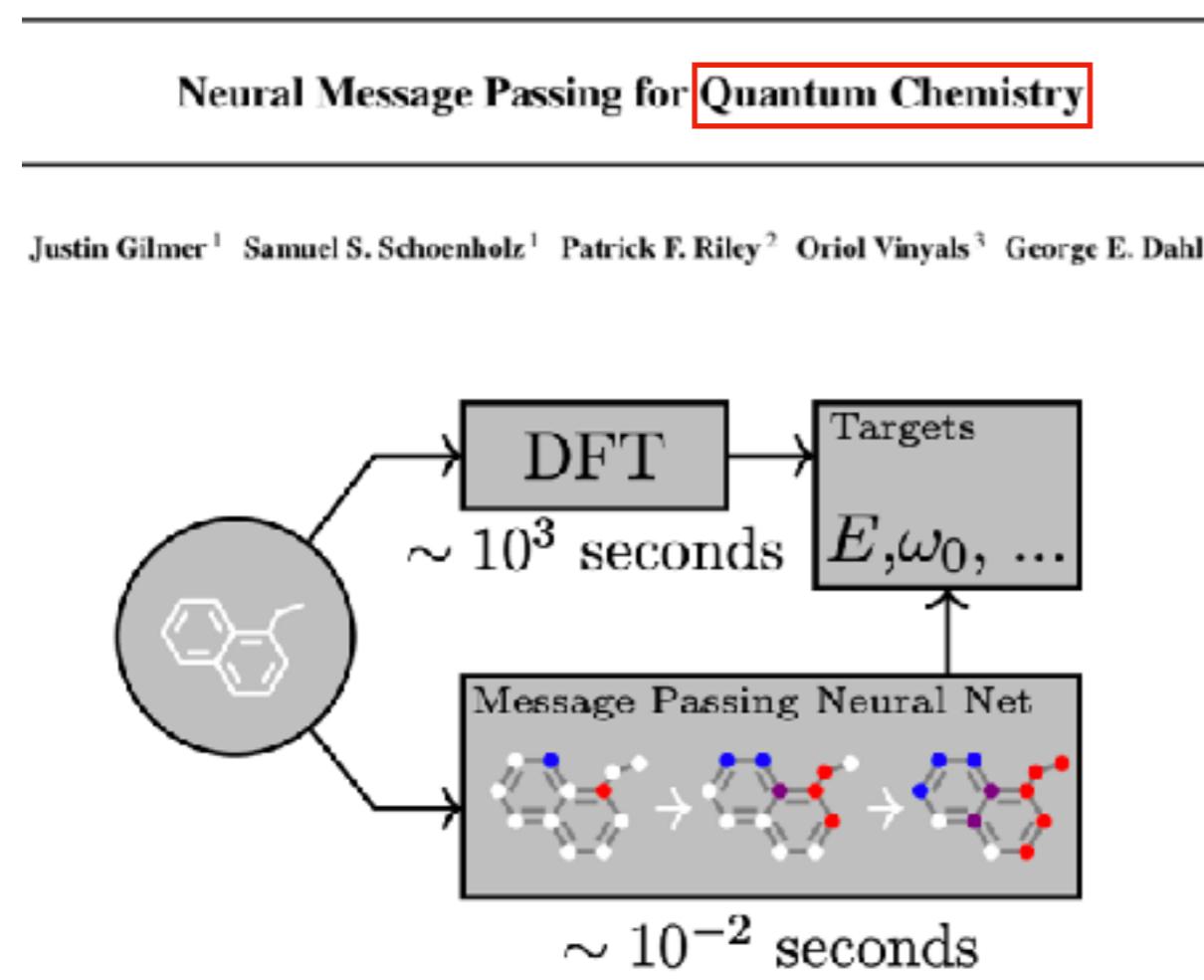


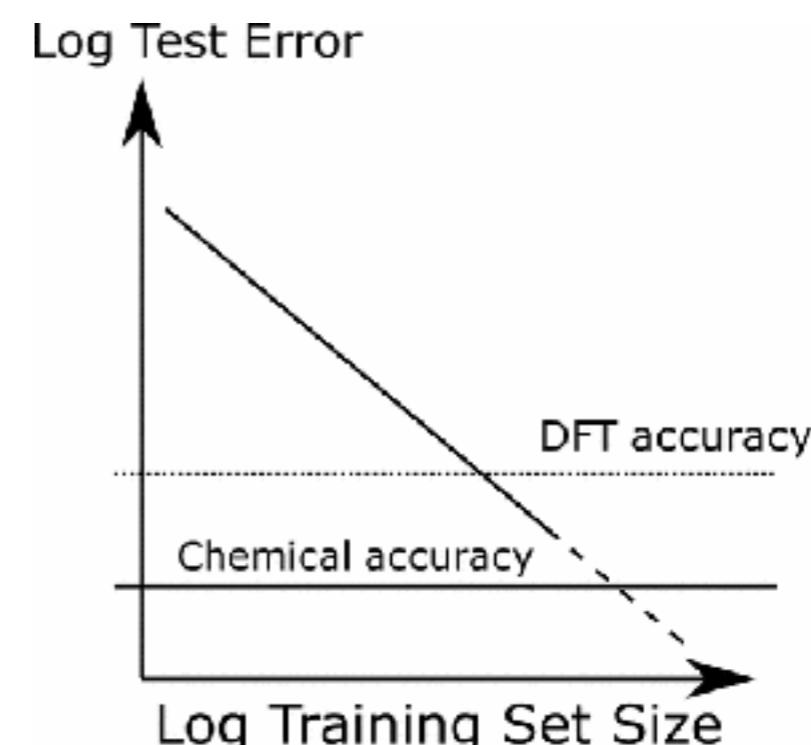
Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

JCTC 2017 <https://doi.org/10.1021/acs.jctc.7b00577>



## Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error

Felix A. Faber,<sup>†</sup> Luke Hutchison,<sup>‡</sup> Bing Huang,<sup>†</sup> Justin Gilmer,<sup>‡</sup> Samuel S. Schoenholz,<sup>‡</sup> George E. Dahl,<sup>‡</sup> Oriol Vinyals,<sup>†</sup> Steven Kearns,<sup>†</sup> Patrick F. Riley,<sup>†</sup> and O. Anatole von Lilienfeld<sup>†,‡,¶</sup>



# Use Case 2: Quantum chemistry

## オリジナルのECFP原子不变量

- the number of immediate neighbors who are “heavy” (non-hydrogen) atoms
- the valence minus the number of hydrogens
- the atomic number
- the atomic mass
- the atomic charge
- the number of attached hydrogens
- whether the atom is contained in at least one ring

Daylight  
原子不变量

## オリジナルのFCFP原子不变量

- hydrogen-bond acceptor or not?
- hydrogen-bond donor or not?
- negatively ionizable or not?
- positively ionizable or not?
- aromatic or not?
- halogen or not?

Rogers and Hahn, *JCIM* (2005) <https://doi.org/10.1021/ci100050t>

## MPNNによる量子化学計算近似で用いられた頂点・辺特徴

Table 1. Atom Features for the MG Representation<sup>a</sup>

feature	description
atom type	H, C, N, O, F (one-hot)
chirality	R or S (one-hot or null)
formal charge	integer electronic charge
ring sizes	for each ring size (3–8), the number of rings that include this atom
hybridization	$sp$ , $sp^2$ , or $sp^3$ (one-hot or null)
hydrogen bonding	whether this atom is a hydrogen bond donor and/or acceptor (binary values)
aromaticity	whether this atom is part of an aromatic system

Table 2. Atom Pair Features for the MG Representation<sup>a</sup>

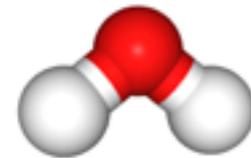
feature	description
bond type	single, double, triple, or aromatic (one-hot or null)
graph distance	for each distance (1–7), whether the shortest path between the atoms in the pair is less than or equal to that number of bonds (binary values)
same ring	whether the atoms in the pair are in the same ring
spatial distance	the Euclidean distance between the two atoms

連続量ラベル

Faber et al, *JCTC* (2017) <https://doi.org/10.1021/acs.jctc.7b00577>

# SchNet

input molecule H<sub>2</sub>O



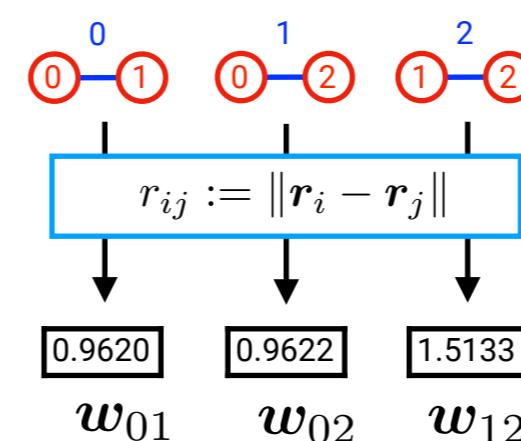
gdb\_3

	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

atom features

$$\begin{array}{ccc} \textcircled{0} & \textcircled{1} & \textcircled{2} \\ 8 & 1 & 1 \\ \boldsymbol{x}_0 & \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{array}$$

bond features

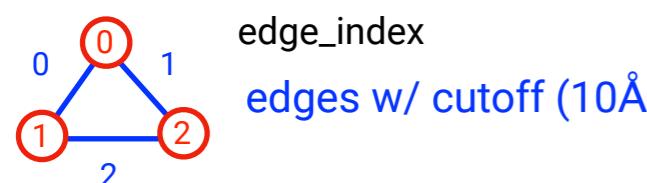


**SchNet (Schütt et al, 2017)**

Message Passing with residual connections

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \psi \left( \sum_{j \in \mathcal{N}_i} \phi(\boldsymbol{x}_j) \odot \omega_{ij} \right)$$

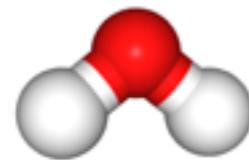
graph (SchNet)



$$\begin{aligned} \mathbf{r}_0 &= [-0.0344, 0.9775, 0.0076] & Z_0 &= 8 \\ \mathbf{r}_1 &= [0.0648, 0.0206, 0.0015] & Z_1 &= 1 \\ \mathbf{r}_2 &= [0.8718, 1.3008, 0.0007] & Z_2 &= 1 \end{aligned}$$

# SchNet

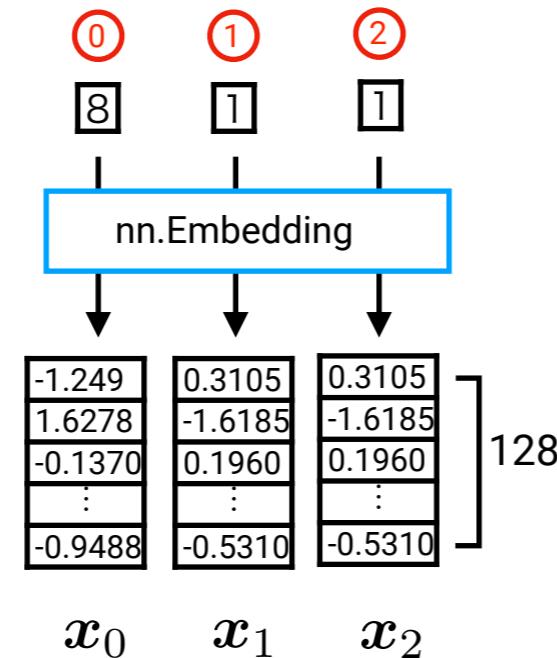
input molecule H<sub>2</sub>O



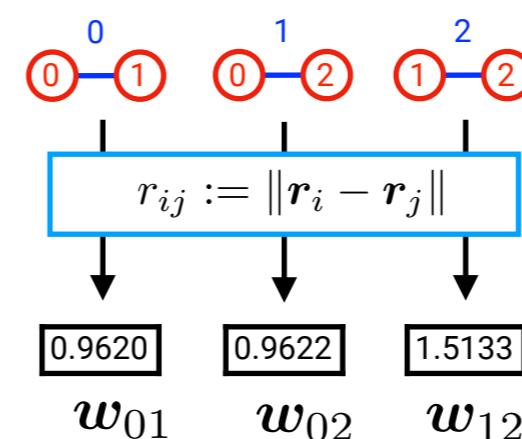
gdb\_3

	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

atom features



bond features

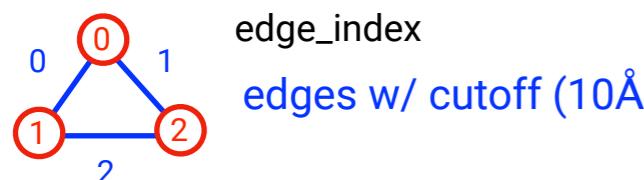


**SchNet (Schütt et al, 2017)**

Message Passing with residual connections

$$\mathbf{x}_i \leftarrow \mathbf{x}_i + \psi \left( \sum_{j \in \mathcal{N}_i} \phi(\mathbf{x}_j) \odot \omega_{ij} \right)$$

graph (SchNet)



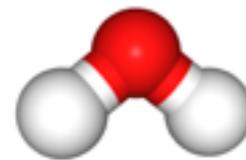
edge\_index

edges w/ cutoff (10Å)

$$\begin{aligned} \mathbf{r}_0 &= [-0.0344, 0.9775, 0.0076] & Z_0 &= 8 \\ \mathbf{r}_1 &= [0.0648, 0.0206, 0.0015] & Z_1 &= 1 \\ \mathbf{r}_2 &= [0.8718, 1.3008, 0.0007] & Z_2 &= 1 \end{aligned}$$

# SchNet

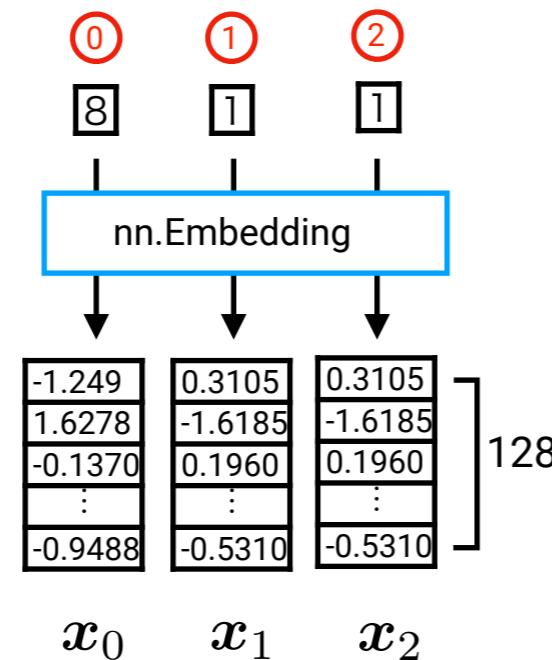
input molecule H<sub>2</sub>O



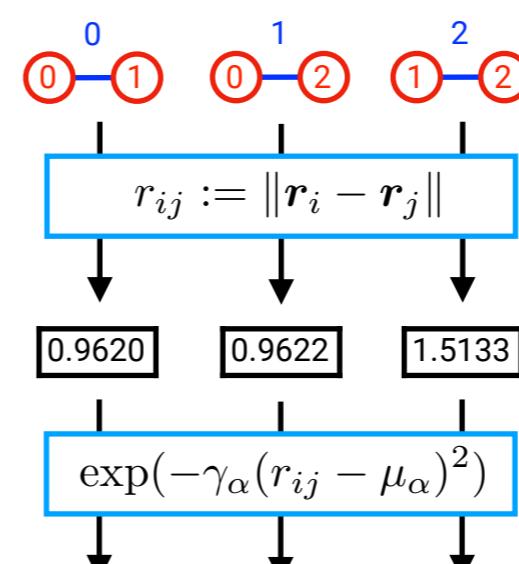
gdb\_3

	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

atom features



bond features



**SchNet (Schütt et al, 2017)**

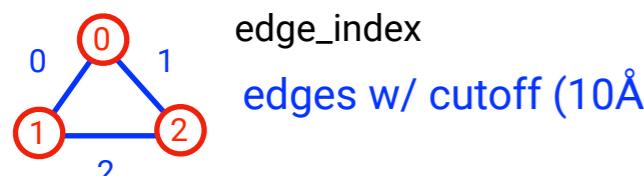
Message Passing with residual connections

$$\mathbf{x}_i \leftarrow \mathbf{x}_i + \psi \left( \sum_{j \in \mathcal{N}_i} \phi(\mathbf{x}_j) \odot \omega_{ij} \right)$$

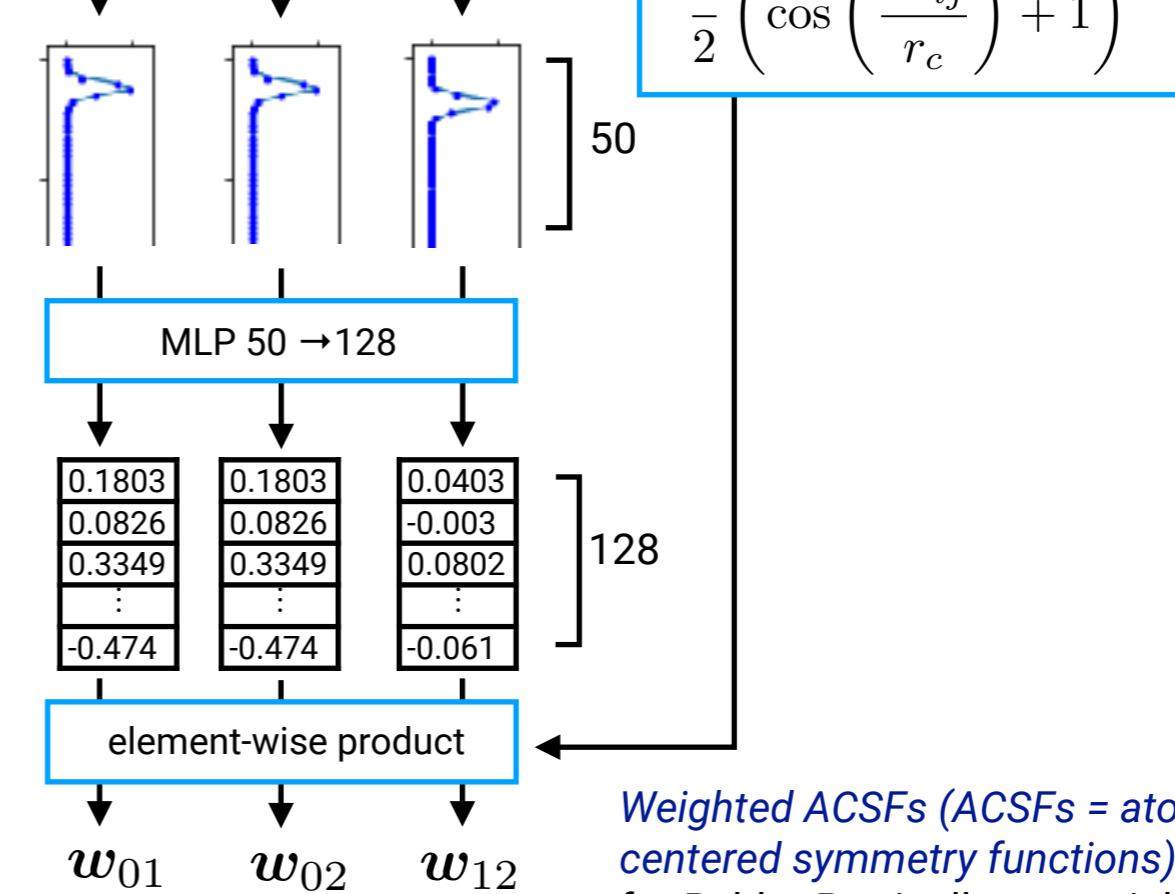
cutoff function

$$\frac{1}{2} \left( \cos \left( \frac{\pi r_{ij}}{r_c} \right) + 1 \right)$$

graph (SchNet)

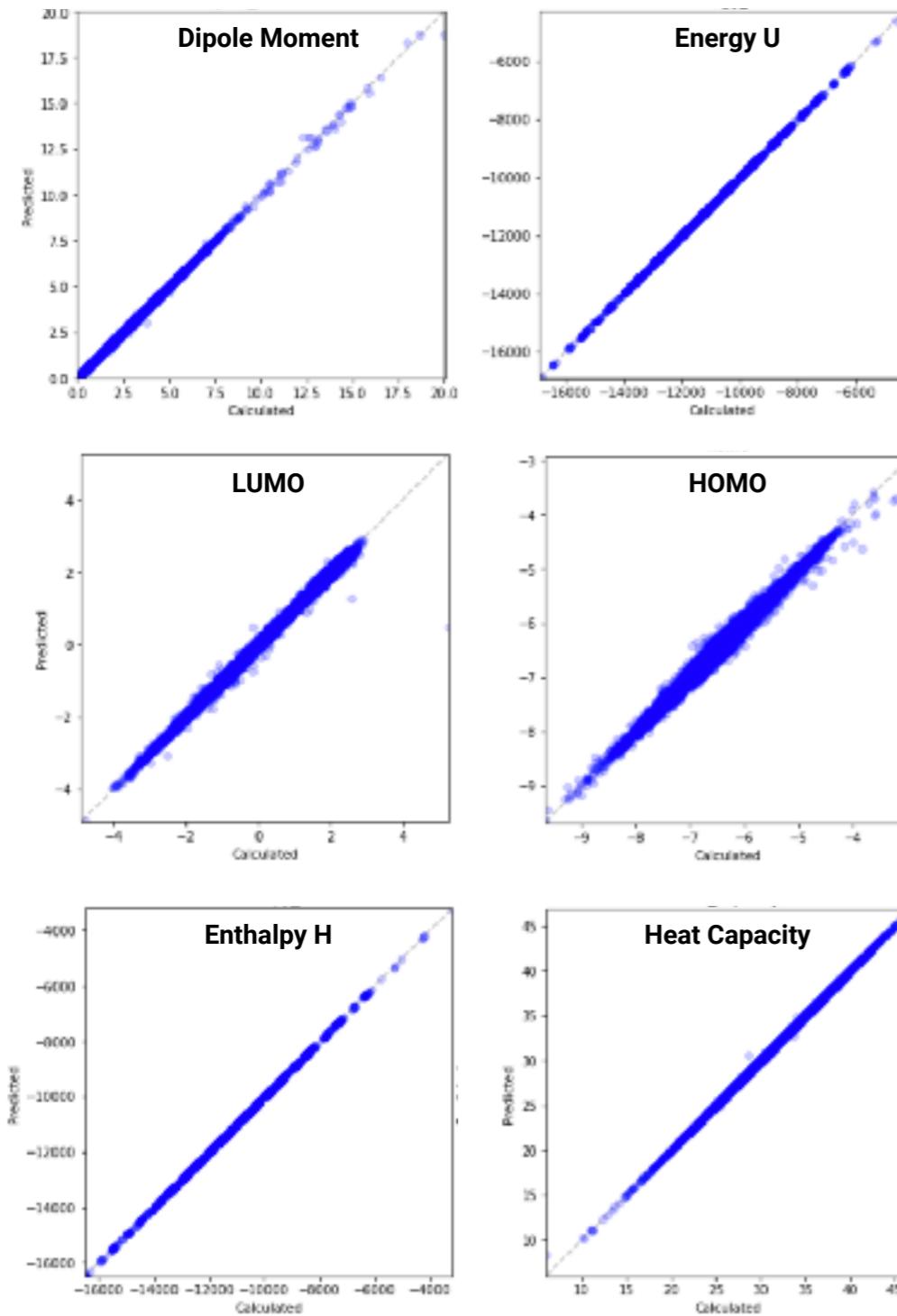


$$\begin{aligned} \mathbf{r}_0 &= [-0.0344, 0.9775, 0.0076] & Z_0 &= 8 \\ \mathbf{r}_1 &= [0.0648, 0.0206, 0.0015] & Z_1 &= 1 \\ \mathbf{r}_2 &= [0.8718, 1.3008, 0.0007] & Z_2 &= 1 \end{aligned}$$

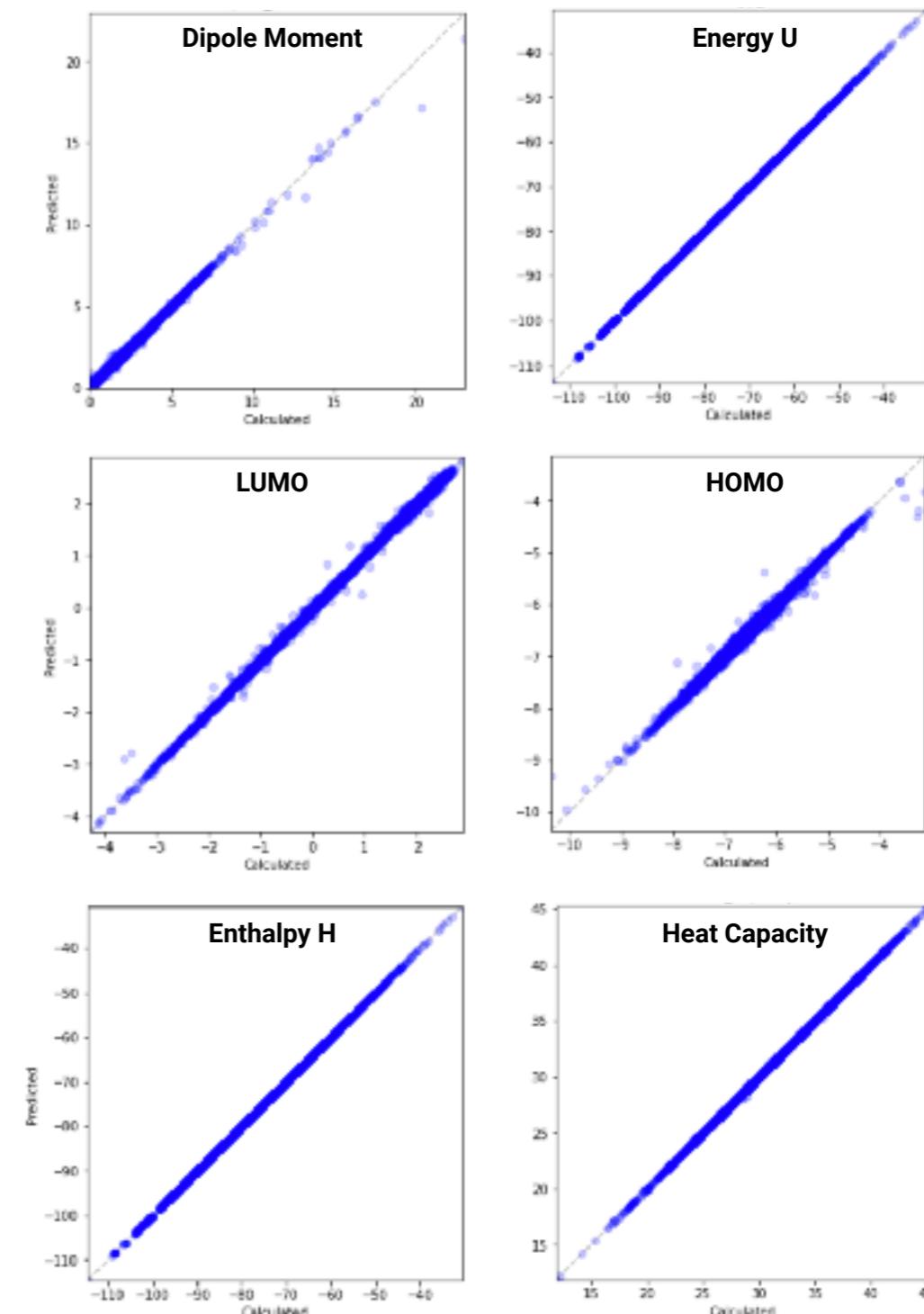


# Use Case 2: Quantum chemistry

pred vs true for **SchNet** (Schütt et al, 2017)

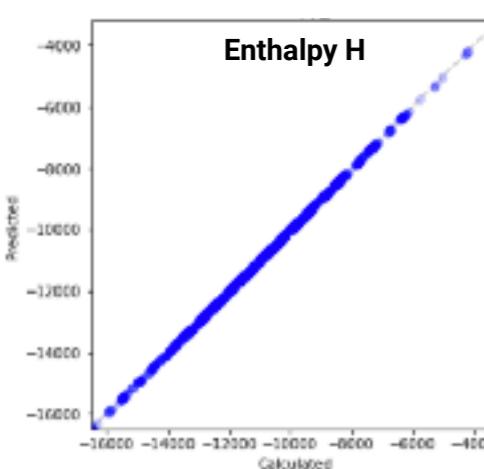
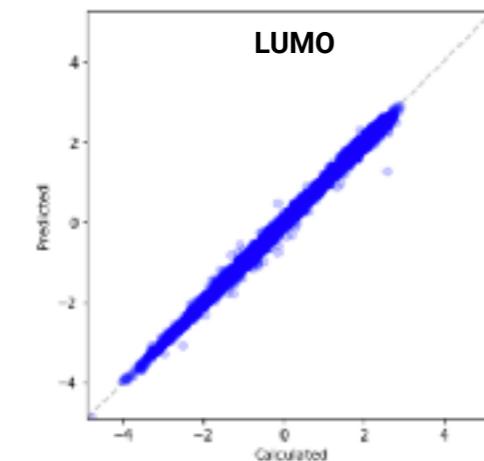
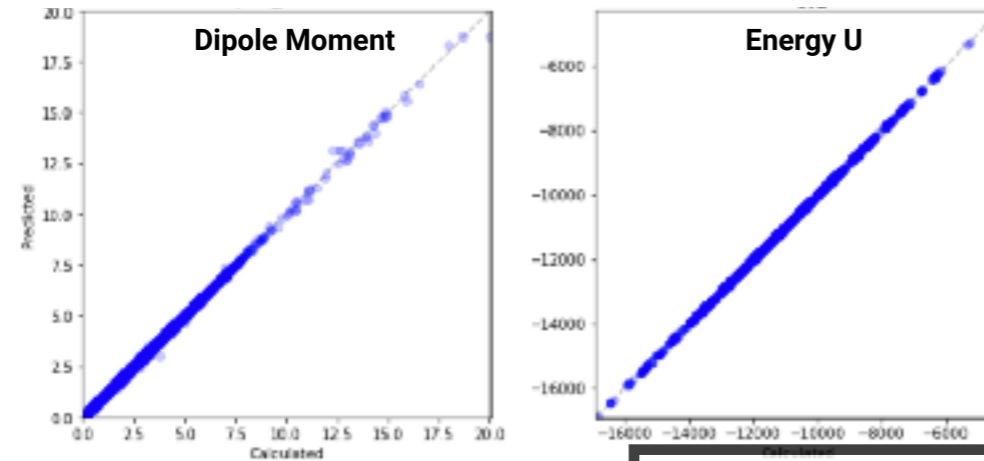


pred vs true for **DimeNet** (Klicpera et al, 2020)

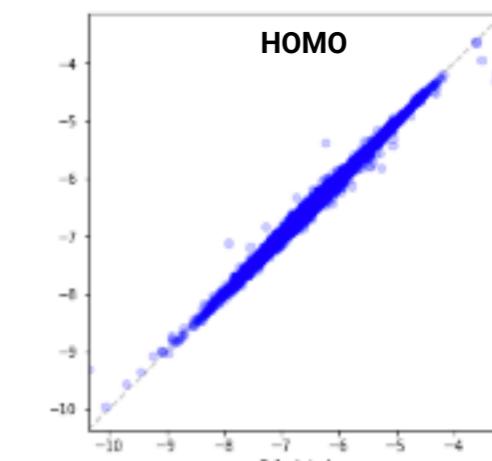
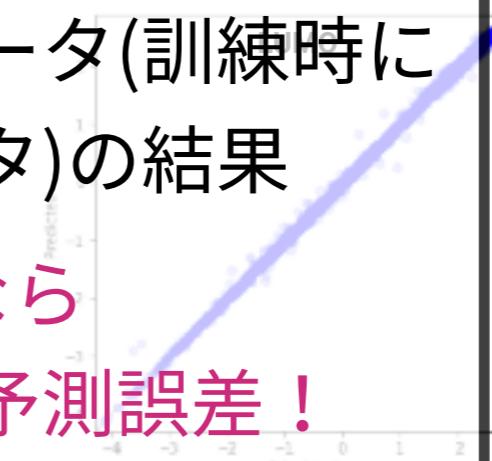
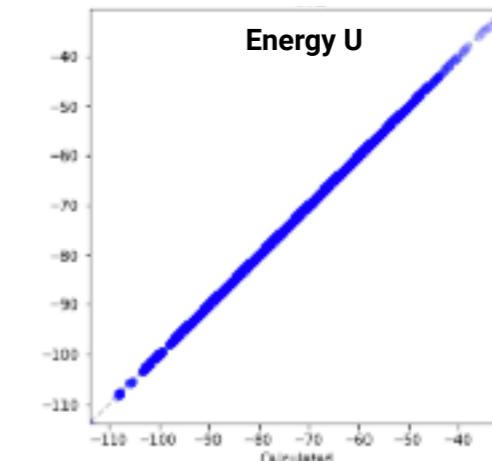
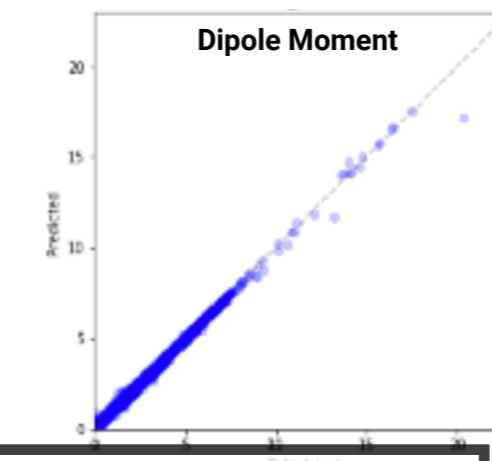


# Use Case 2: Quantum chemistry

pred vs true for **SchNet** (Schütt et al, 2017)

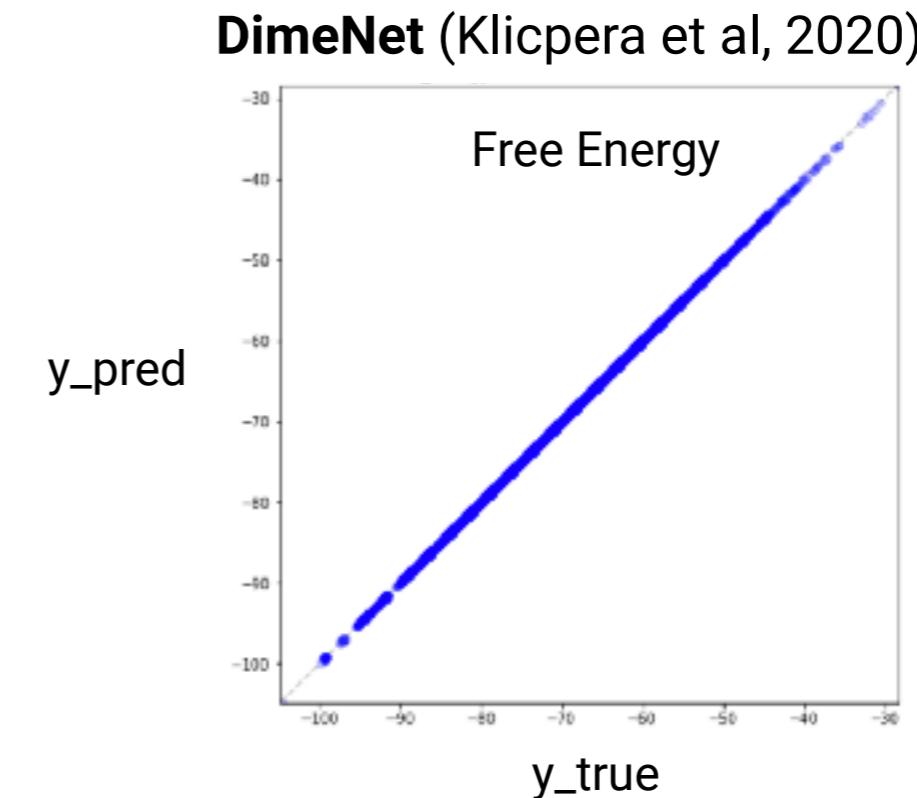
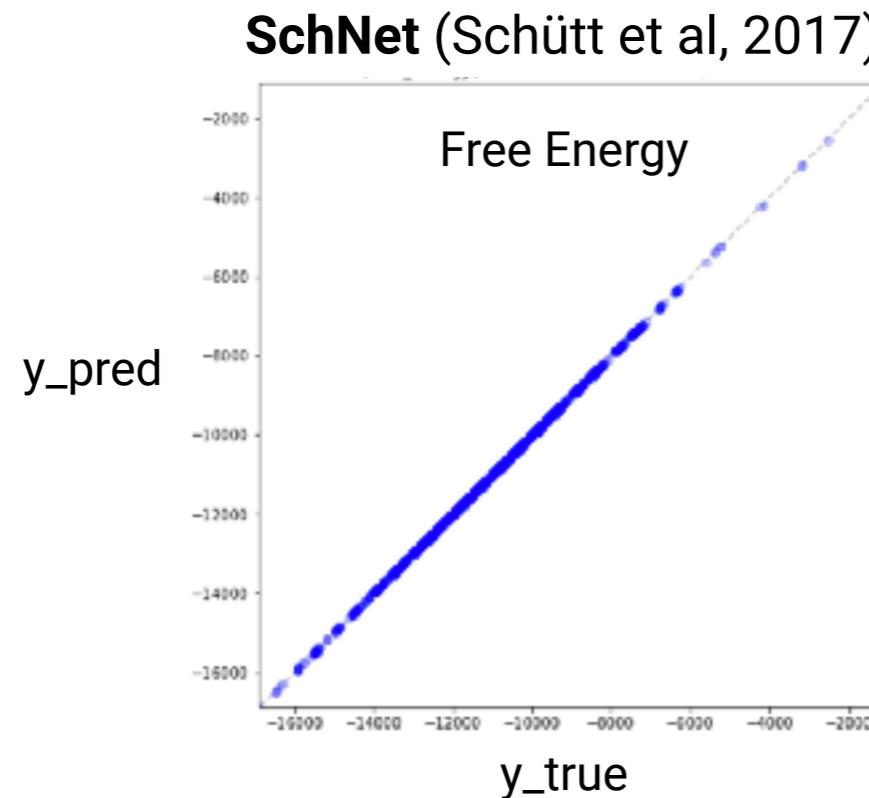


pred vs true for **DimeNet** (Klicpera et al, 2020)

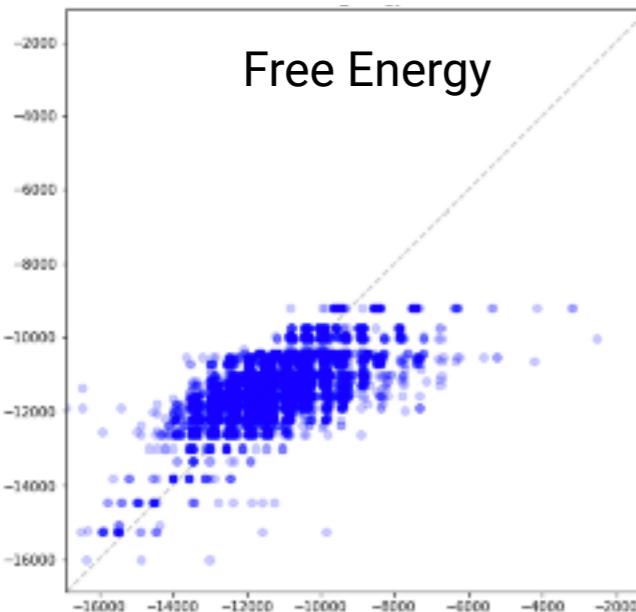


これはテストデータ(訓練時に  
見せてないデータ)の結果  
100,000倍速いなら  
十分許容できる予測誤差!

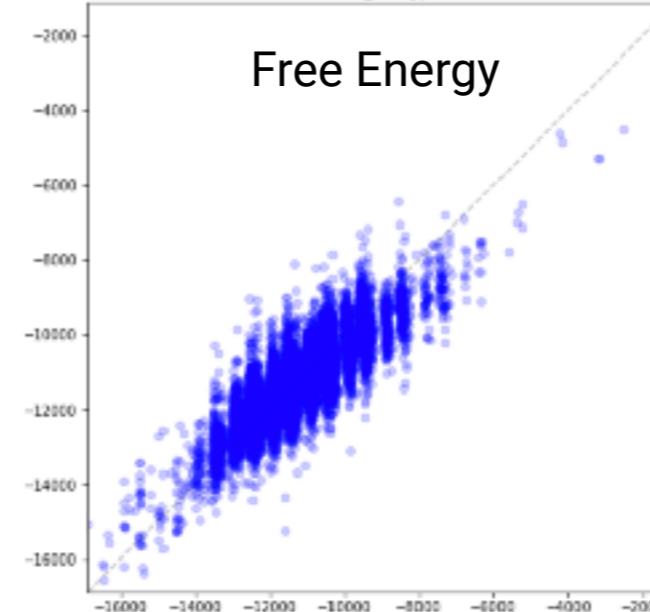
# Use Case 2: Quantum chemistry



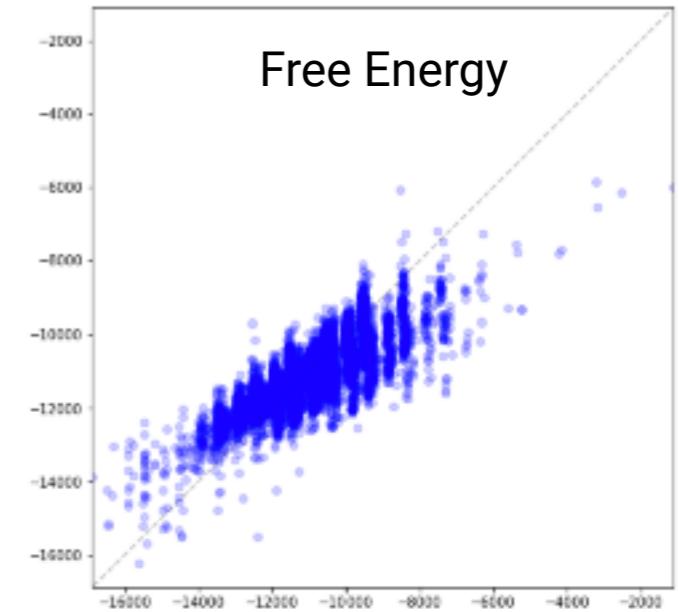
**ExtraTrees w/ ECFP6**  
(without 3D geometry)



**LightGBM w/ ECFP6**  
(without 3D geometry)



**3-Layer MLP w/ ECFP6**  
(without 3D geometry)



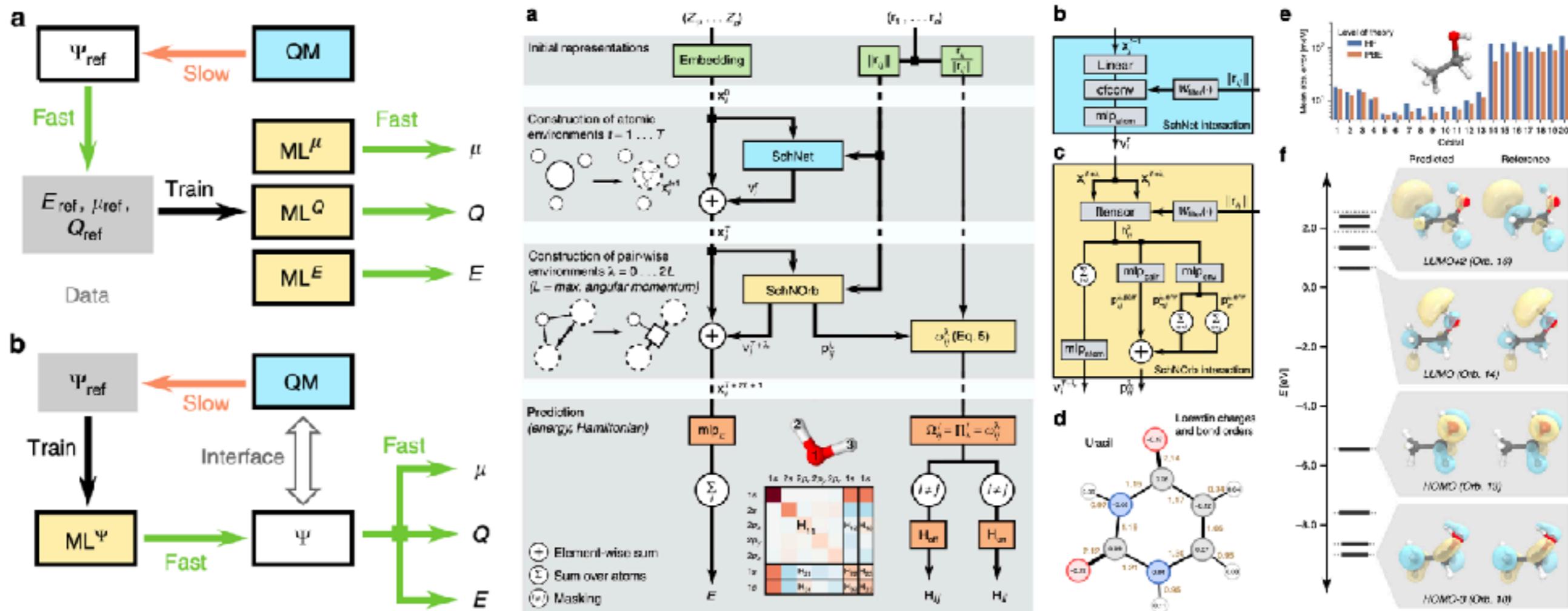
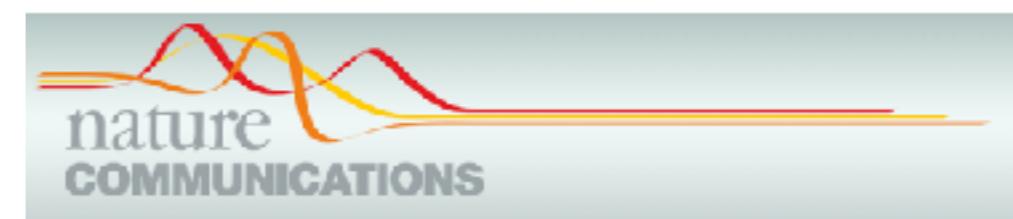
# SchNOrb: 波動関数 자체を機械学習

ARTICLE

<https://doi.org/10.1038/s41467-019-12875-2>

OPEN

## Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions

K.T. Schütt<sup>1</sup>, M. Gastegger<sup>1</sup>, A. Tkatchenko<sup>2\*</sup>, K.-R. Müller<sup>1,3,4\*</sup> & R.J. Maurer<sup>5\*</sup>

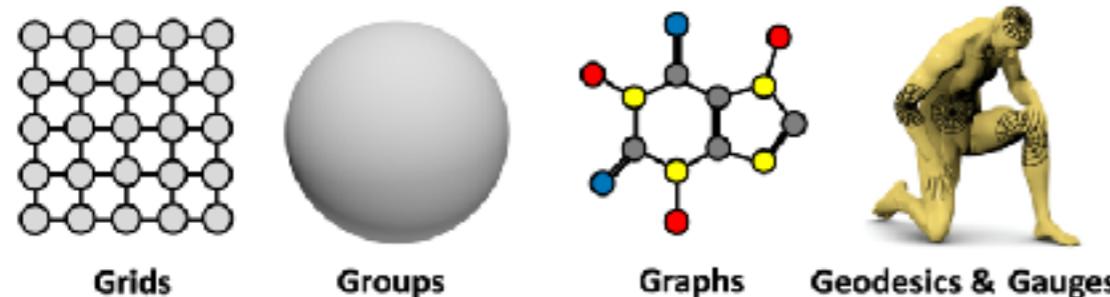
# GNNs for Geometric Deep Learning

<https://arxiv.org/abs/2104.13478>

[Submitted on 27 Apr 2021 ([v1](#)), last revised 2 May 2021 (this version, v2)]

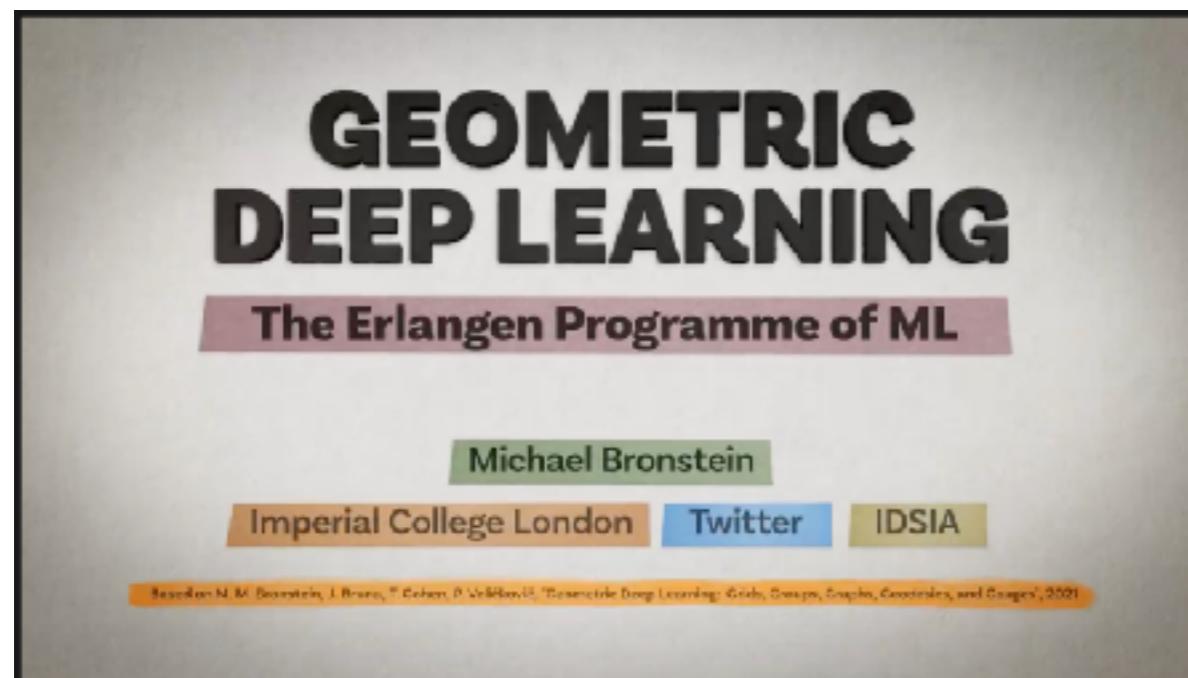
## Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges

Michael M. Bronstein, Joan Bruna, Taco Cohen, Petar Veličković



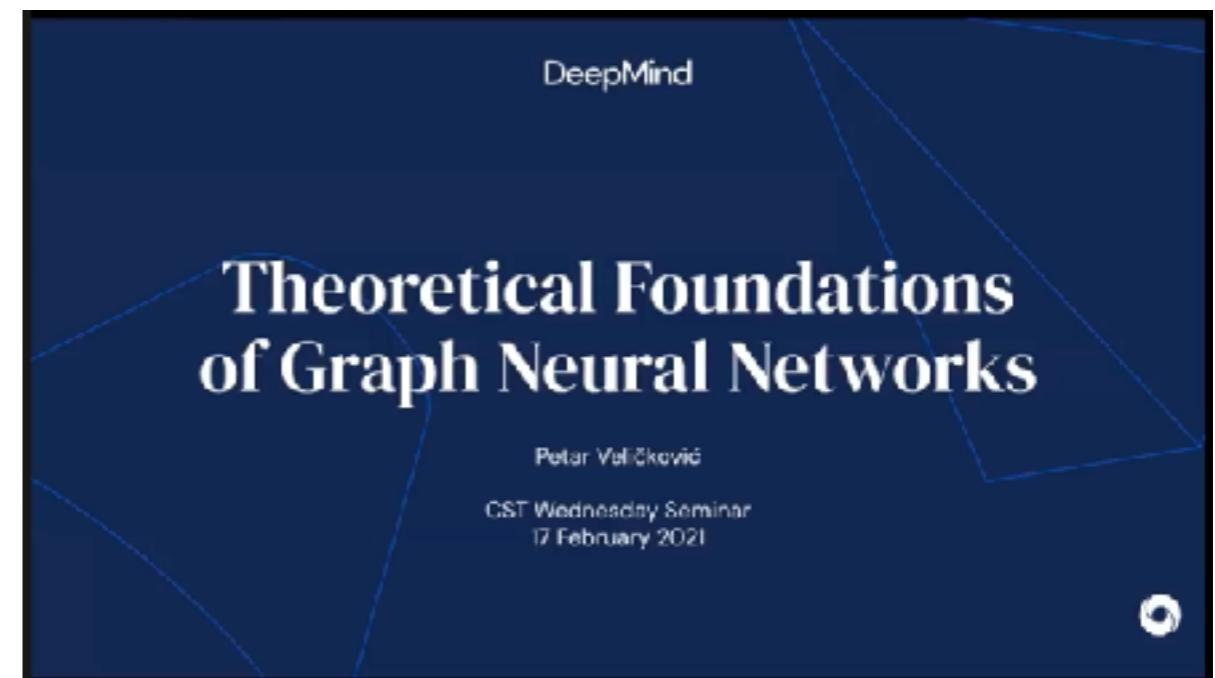
GNNは幅広い幾何構造を統一的に扱える枠組み  
(機械学習のエルランゲン・プログラム!?)  
**5Gs: Grids, Groups, Graphs, Geodesics/Gauges**

ICLR 2021 Keynote (Michael Bronstein)



<https://youtu.be/w6Pw4MOzMuo>

Seminar Talk (Petar Veličković)



<https://youtu.be/uF53xsT7mjc>

# ユークリッドの運動群に関する不变性・同変性

原子のxyz座標値をそのまま頂点特徴量にするのは× (代案の例: 原子間距離を辺特徴に)

平行移動や回転でxyzは変わるが例えばその分子のエネルギーは変わらない

**幾何的GNNでは基本的な要請  
(特に量子化学計算近似の場合)**

- ユークリッド群  $E(3)$  : 3Dの並進・回転対称性
- 特殊ユークリッド群  $SE(3)$  : 3Dの並進・回転・鏡像対称性

写像  $f : X \rightarrow Y$  が変換  $g \in G$  に関して

**不变 (invariant)**  $f(g \cdot x) = f(x)$  変換してもしないときと変わらない

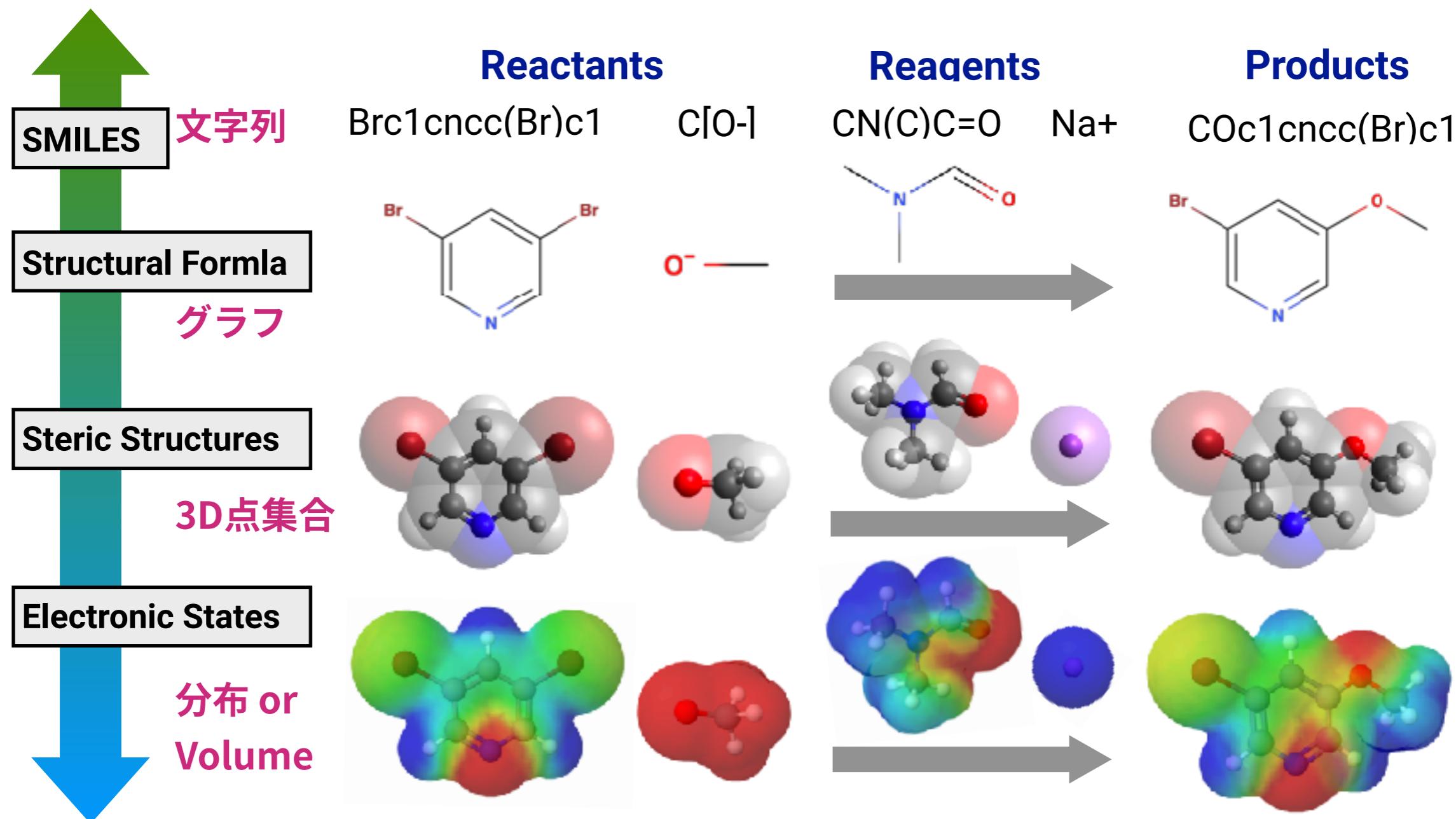
**同変 (equivariant)**  $f(g \cdot x) = g \cdot f(x)$  変換してから写像しても写像してから変換しても変わらない

**頂点や辺の特徴量やGNN(=写像)のデザインで実現する**

E(3)不变	Schütt et al, <a href="https://arxiv.org/abs/1706.08566">SchNet</a> . (2017) <a href="https://arxiv.org/abs/1706.08566">https://arxiv.org/abs/1706.08566</a>
	Unke et al, <a href="https://arxiv.org/abs/1902.08408">PhysNet</a> . (2019) <a href="https://arxiv.org/abs/1902.08408">https://arxiv.org/abs/1902.08408</a>
	Klicpera et al, <a href="https://arxiv.org/abs/2011.14115">DimeNet++</a> . (2020) <a href="https://arxiv.org/abs/2011.14115">https://arxiv.org/abs/2011.14115</a>
SE(3)同変	Anderson et al, <a href="https://arxiv.org/abs/1906.04015">Cormorant</a> . (2019) <a href="https://arxiv.org/abs/1906.04015">https://arxiv.org/abs/1906.04015</a>
	Fuchs et al, <a href="https://arxiv.org/abs/2006.10503">SE(3)-Transformers</a> . (2021) <a href="https://arxiv.org/abs/2006.10503">https://arxiv.org/abs/2006.10503</a>
E(3)同変	Thomas et al, <a href="https://arxiv.org/abs/1802.08219">Tensor Field Networks</a> . (2018) <a href="https://arxiv.org/abs/1802.08219">https://arxiv.org/abs/1802.08219</a>
	Köhler et al, <a href="https://arxiv.org/abs/2006.02425">Equivariant Flows (Radial Field)</a> . (2020) <a href="https://arxiv.org/abs/2006.02425">https://arxiv.org/abs/2006.02425</a>
	Satorras et al, <a href="https://arxiv.org/abs/2102.09844">E(n) Equivariant Graph Neural Networks</a> . (2021) <a href="https://arxiv.org/abs/2102.09844">https://arxiv.org/abs/2102.09844</a>

# 分子表現の統一的方法論になりえるか？

パターン言語として (化学の教科書・データベースにある知識表現)

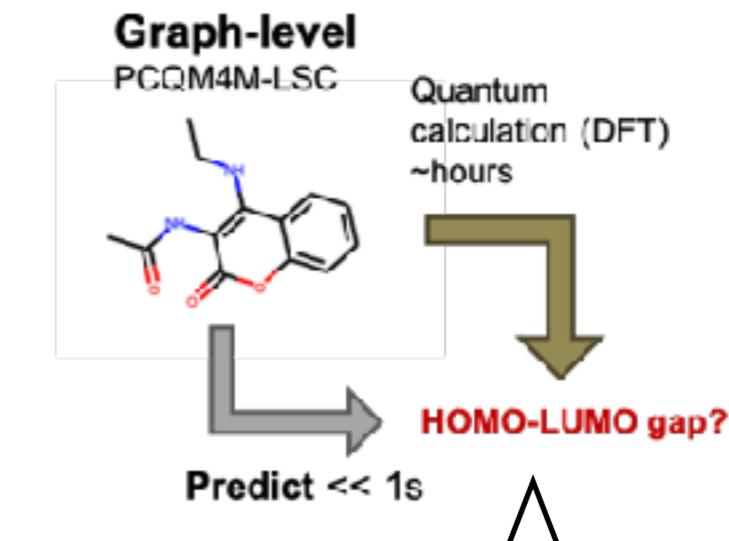
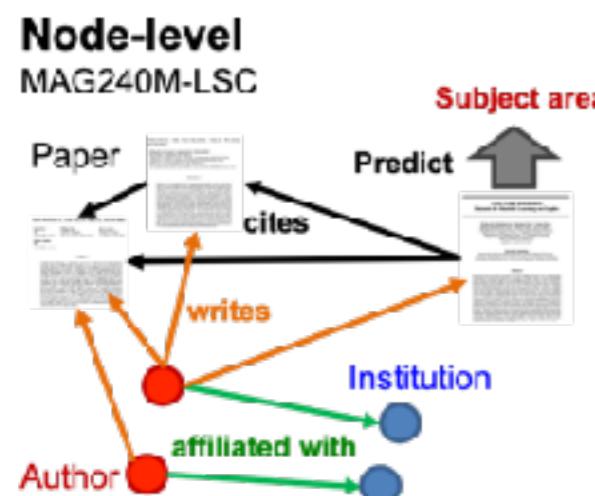


物理的対象として (量子化学に基づく電子状態計算)

# OGB Large-Scale Challenge (KDDCup2021)



<https://ogb.stanford.edu/kddcup2021/>



2Dの分子グラフから量子化学計算(DFT計算)で求めたHOMO-LUMOギャップを予測するタスク  
データセット：PubChemQCから3,803,453グラフ (cf. QM9は133,885グラフ)

**Results:** [https://ogb.stanford.edu/kddcup2021/results/#awardees\\_pcqm4m](https://ogb.stanford.edu/kddcup2021/results/#awardees_pcqm4m)

**1st place:** Test MAE 0.1200 (eV) **10 GNNs (12-Layer Graphomer) + 8 ExpC\*s (5-Layer ExpandingConv)**

**2nd place:** Test MAE 0.1204 (eV) **73 GNNs (11-Layer LiteGEMConv with Self-Supervised Pretraining)**

**3rd place:** Test MAE 0.1205 (eV) **20 GNNs (32-Layer GNN with Noisy Nodes)**

# ICReDD: 化学反応のデザインと発見



中核技術は拠点長・前田 理 教授が開発した量子化学計算に基づく  
化学反応経路自動探索アルゴリズム

S. Maeda, Y. Harabuchi, *Exploring paths of chemical transformations in molecular and periodic systems: An approach utilizing force.*, WIREs Comput. Mol. Sci., 2021, 11, e1538. <https://doi.org/10.1002/wcms.1538>

**AFIR**      **Artificial Force Induced Reaction**  
(人工力誘起反応法)

**GRRM**      **Global Reaction Route Mapping**  
(グローバル反応経路マッピング)

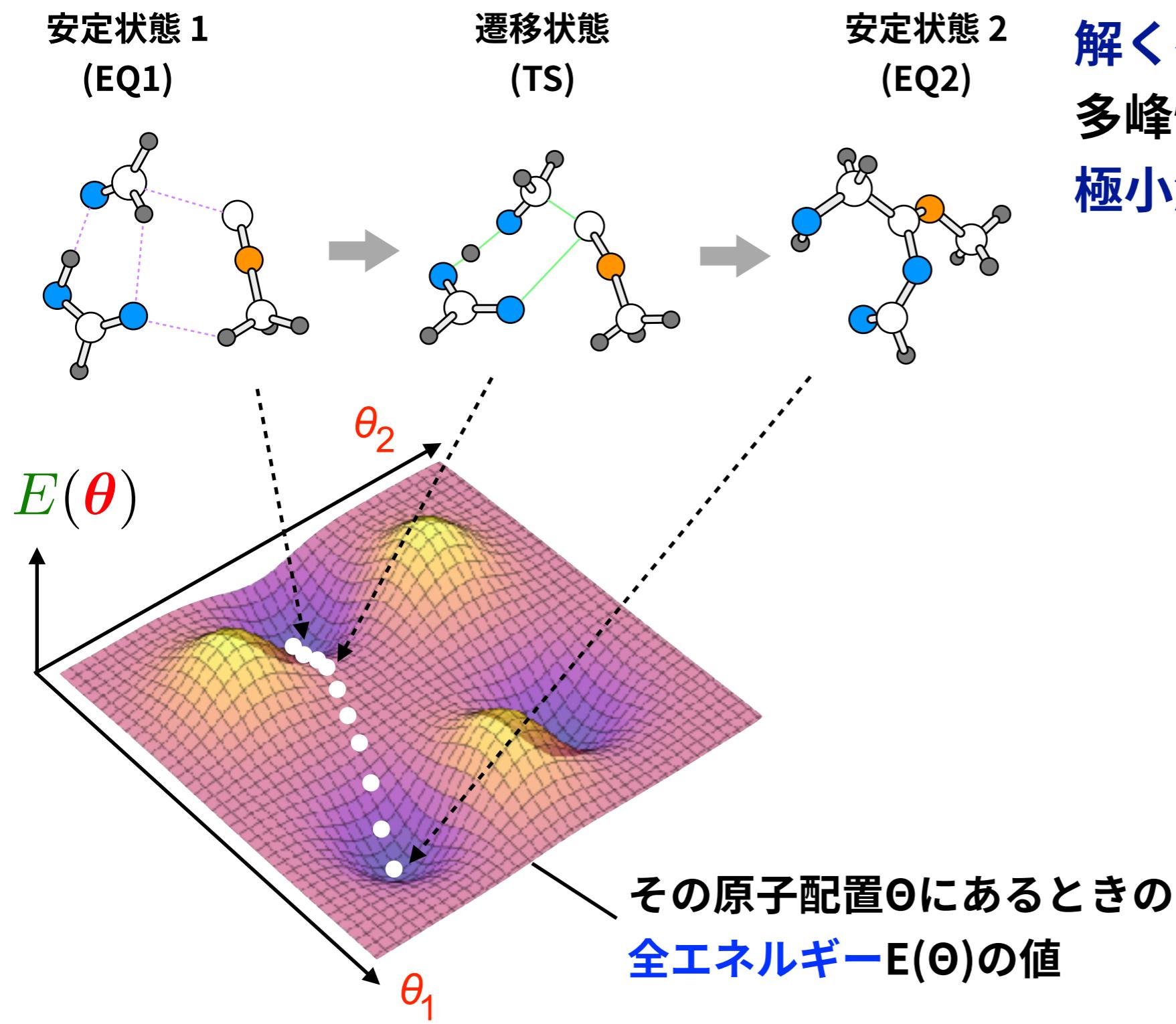


Science as a Service      **GRRM20**  
Software

計算化学ソフトウェア

商用版ソフトウェアの名前にも

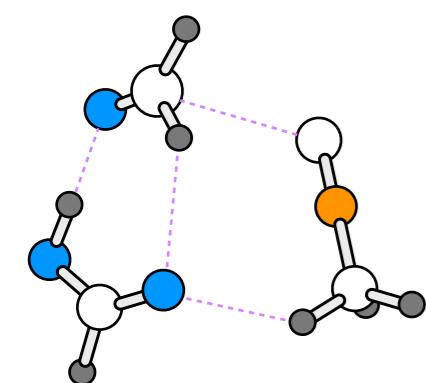
# 情報科学から見た反応経路の自動探索



**解くべき問題：**  
**多峰性の多変数関数の  
極小解と鞍点の「全列挙」**

# 情報科学から見た反応経路の自動探索

安定状態 1  
(EQ1)



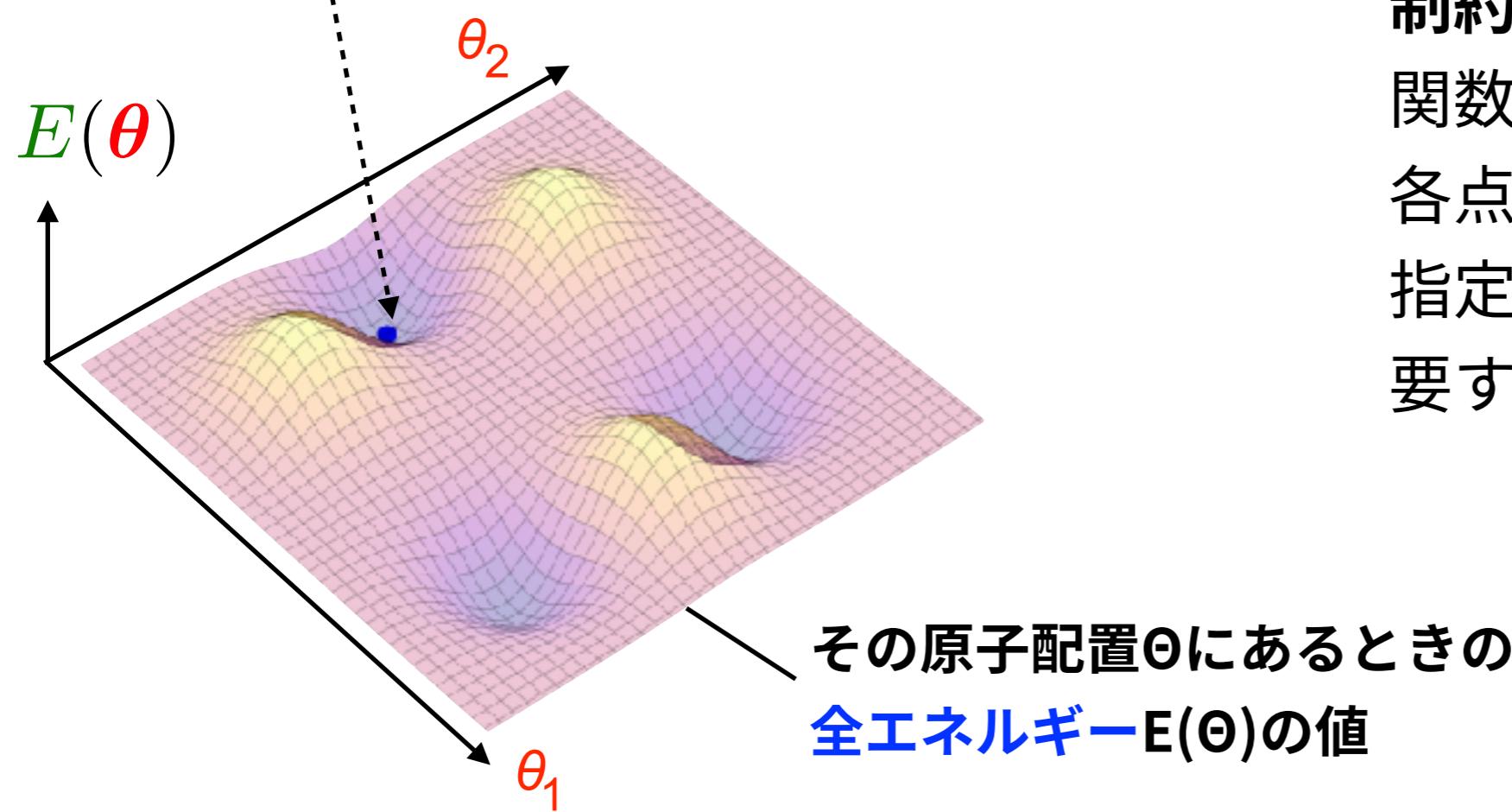
遷移状態  
(TS)

?

安定状態 2  
(EQ2)

?

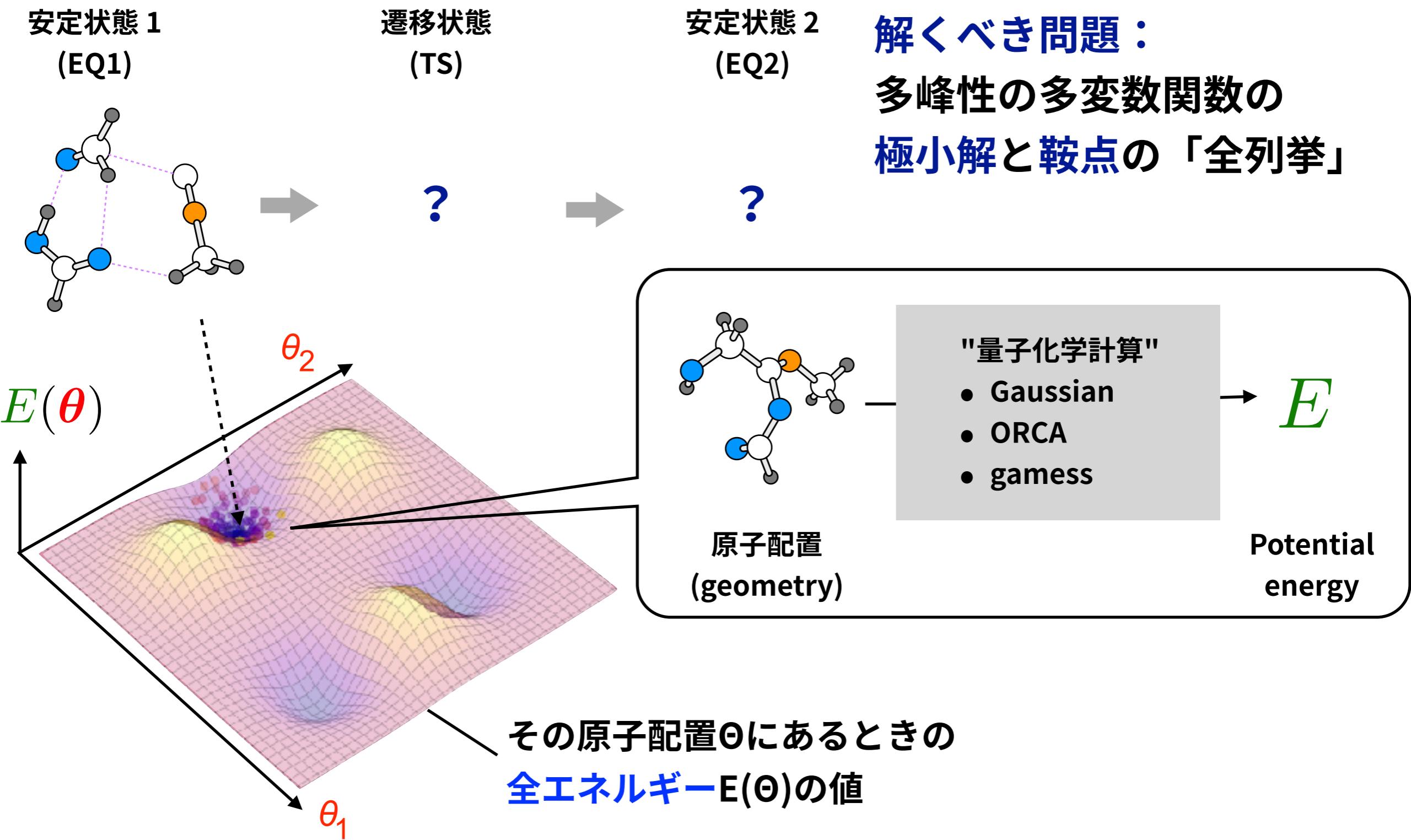
**解くべき問題：**  
多峰性の多変数関数の  
極小解と鞍点の「全列挙」



**制約：**

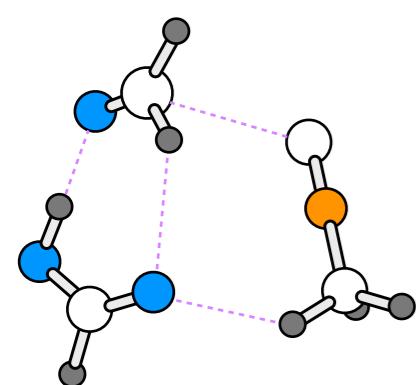
関数形は陽には書けず、  
各点の関数値 $f(x)$ は  
指定した $x$ ごとに計算時間を  
要する問い合わせで得る

# 情報科学から見た反応経路の自動探索



# 情報科学から見た反応経路の自動探索

安定状態 1  
(EQ1)

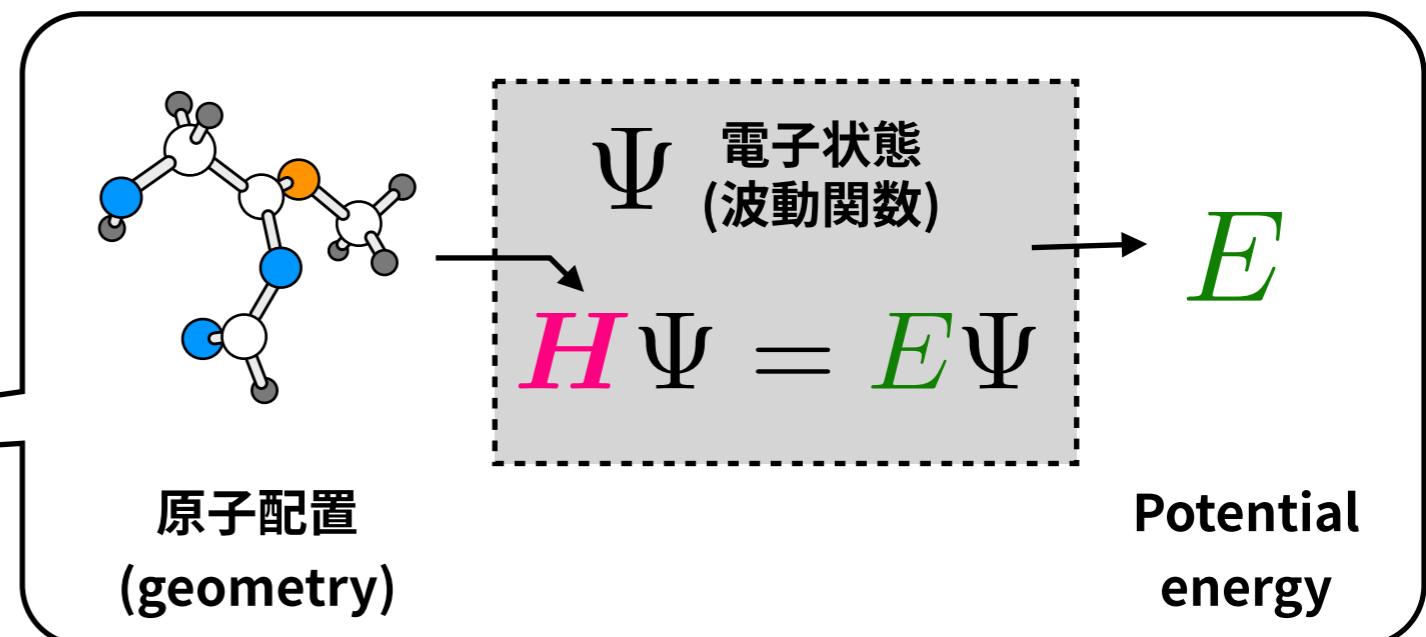
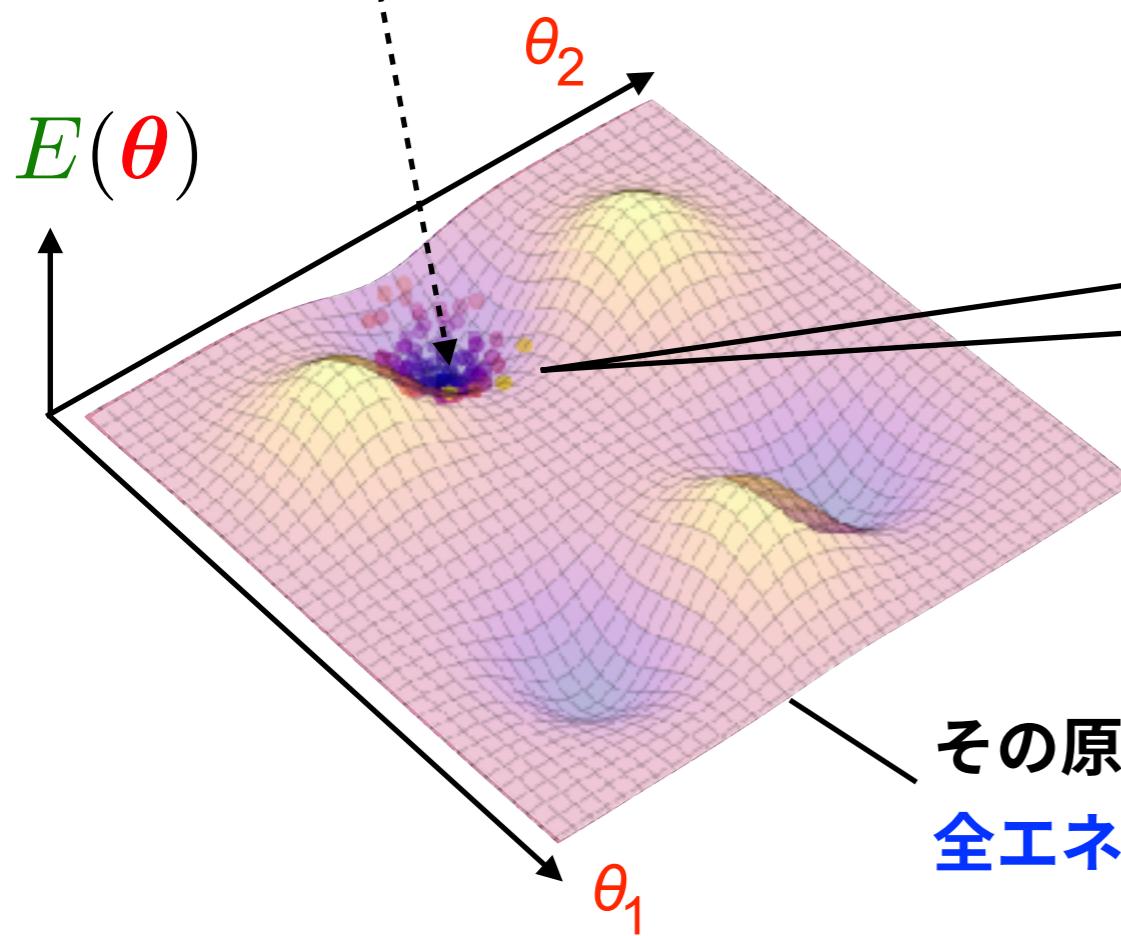


遷移状態  
(TS)

?

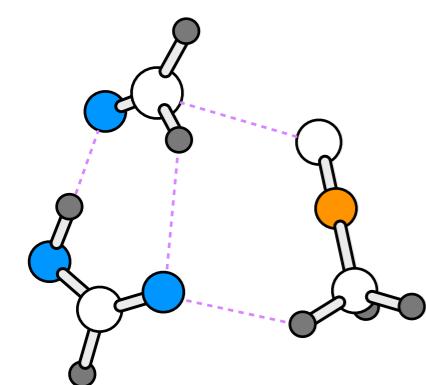
安定状態 2  
(EQ2)

解くべき問題：  
多峰性の多変数関数の  
極小解と鞍点の「全列挙」



# 情報科学から見た反応経路の自動探索

安定状態 1  
(EQ1)

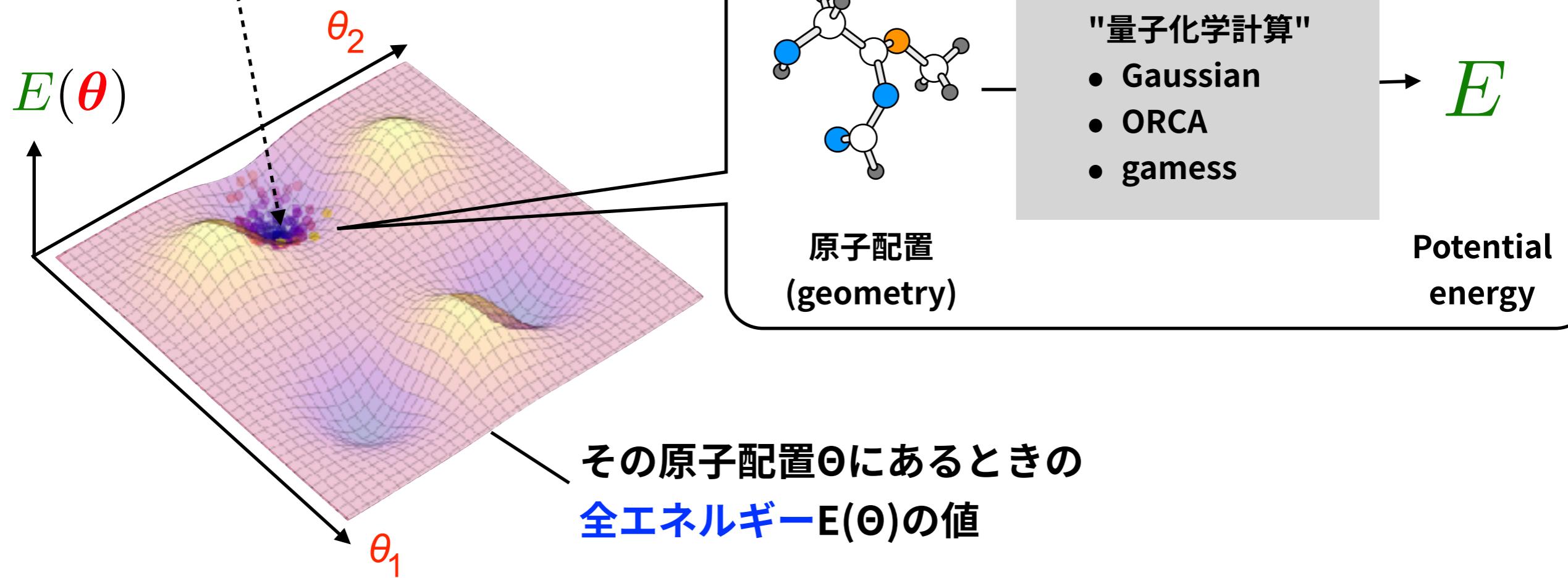


遷移状態  
(TS)

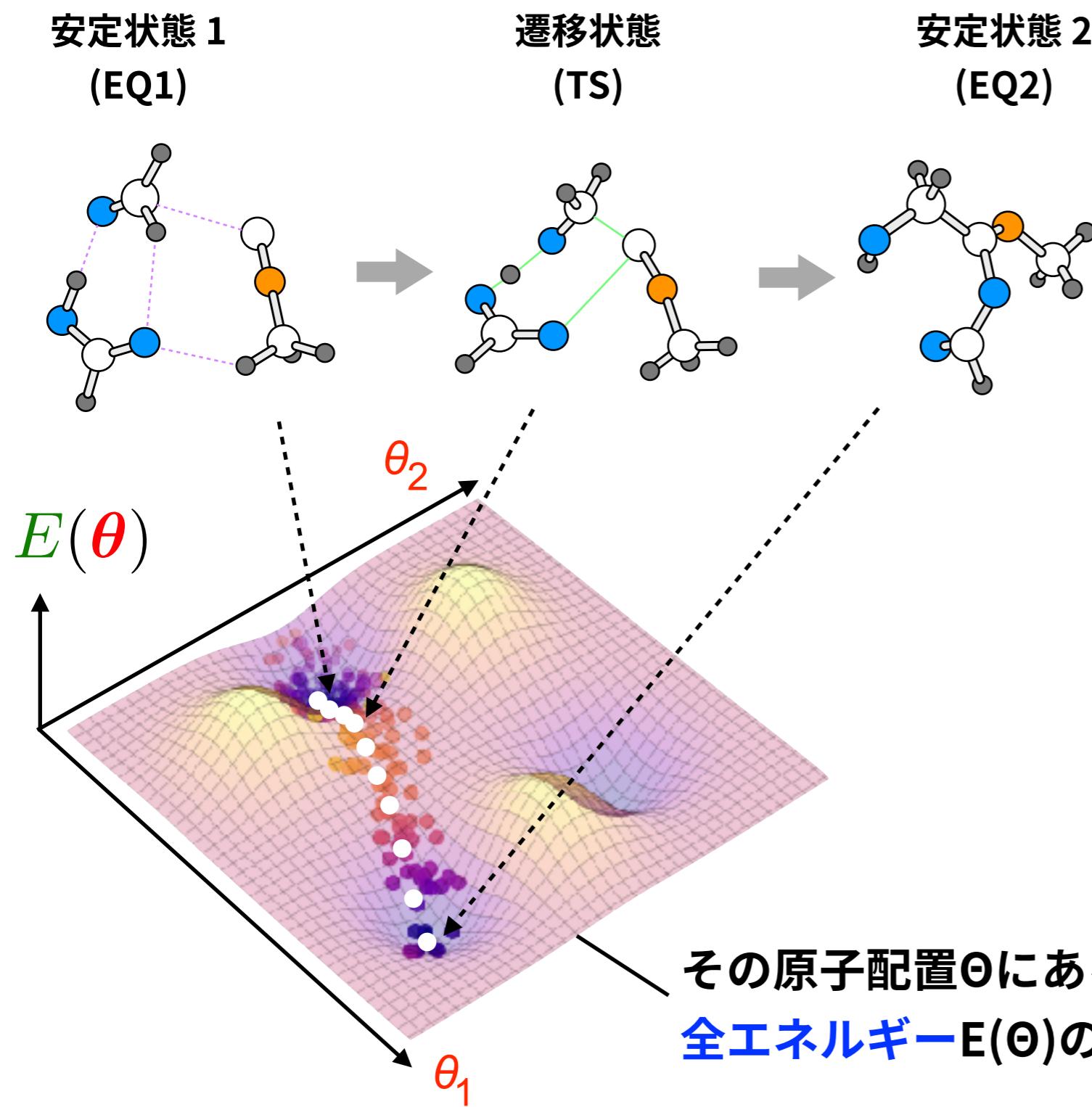
?

安定状態 2  
(EQ2)

解くべき問題：  
多峰性の多変数関数の  
極小解と鞍点の「全列挙」



# 情報科学から見た反応経路の自動探索

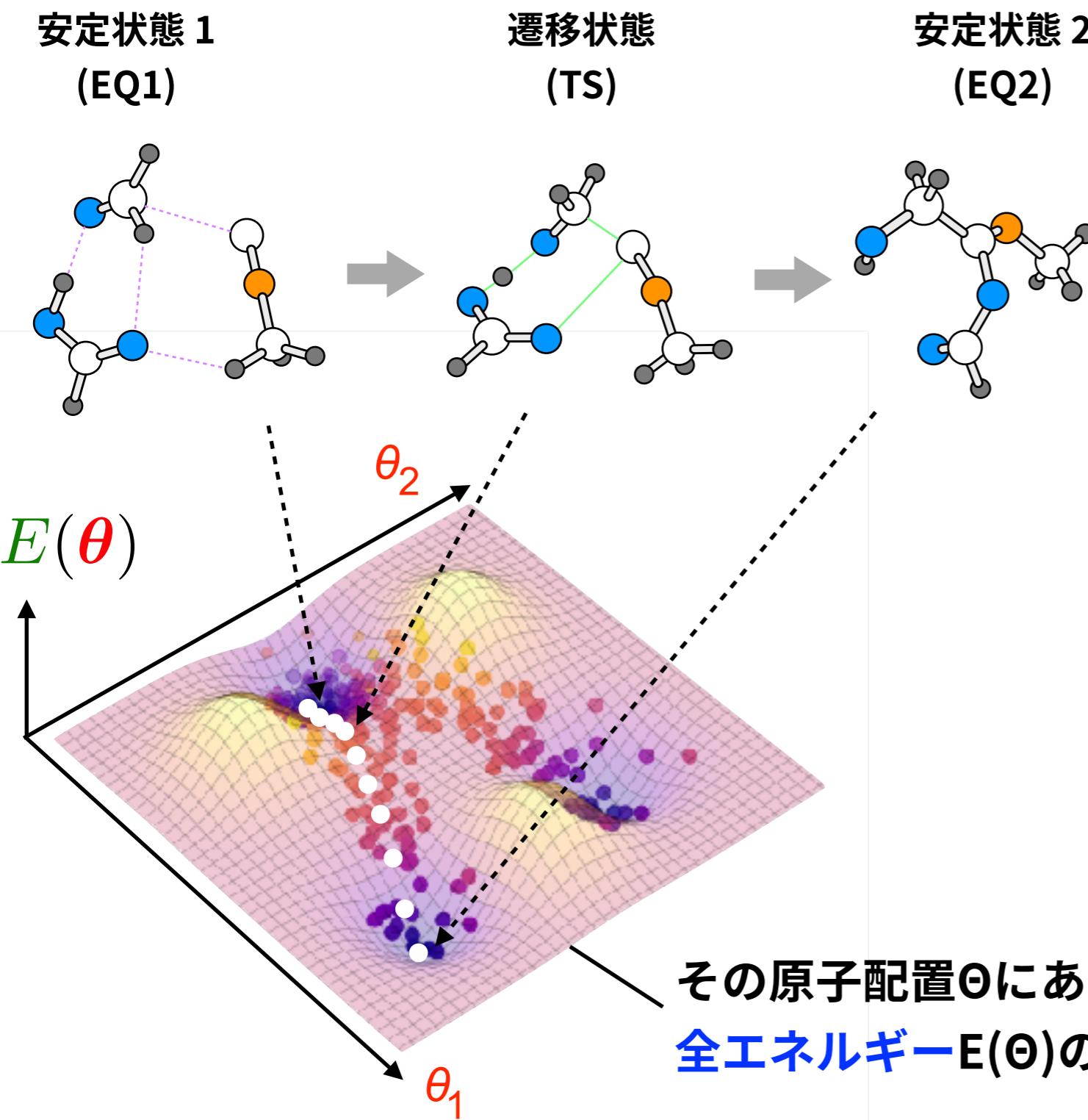


解くべき問題：  
多峰性の多変数関数の  
極小解と鞍点の「全列挙」

そんなの無理では！？

1999年、有名な計算化学の教科書でFrank Jensenは「変数の数が3ないし4を超えると化学反応の遷移状態をすべて求めることは不可能である」と述べた。

# 情報科学から見た反応経路の自動探索

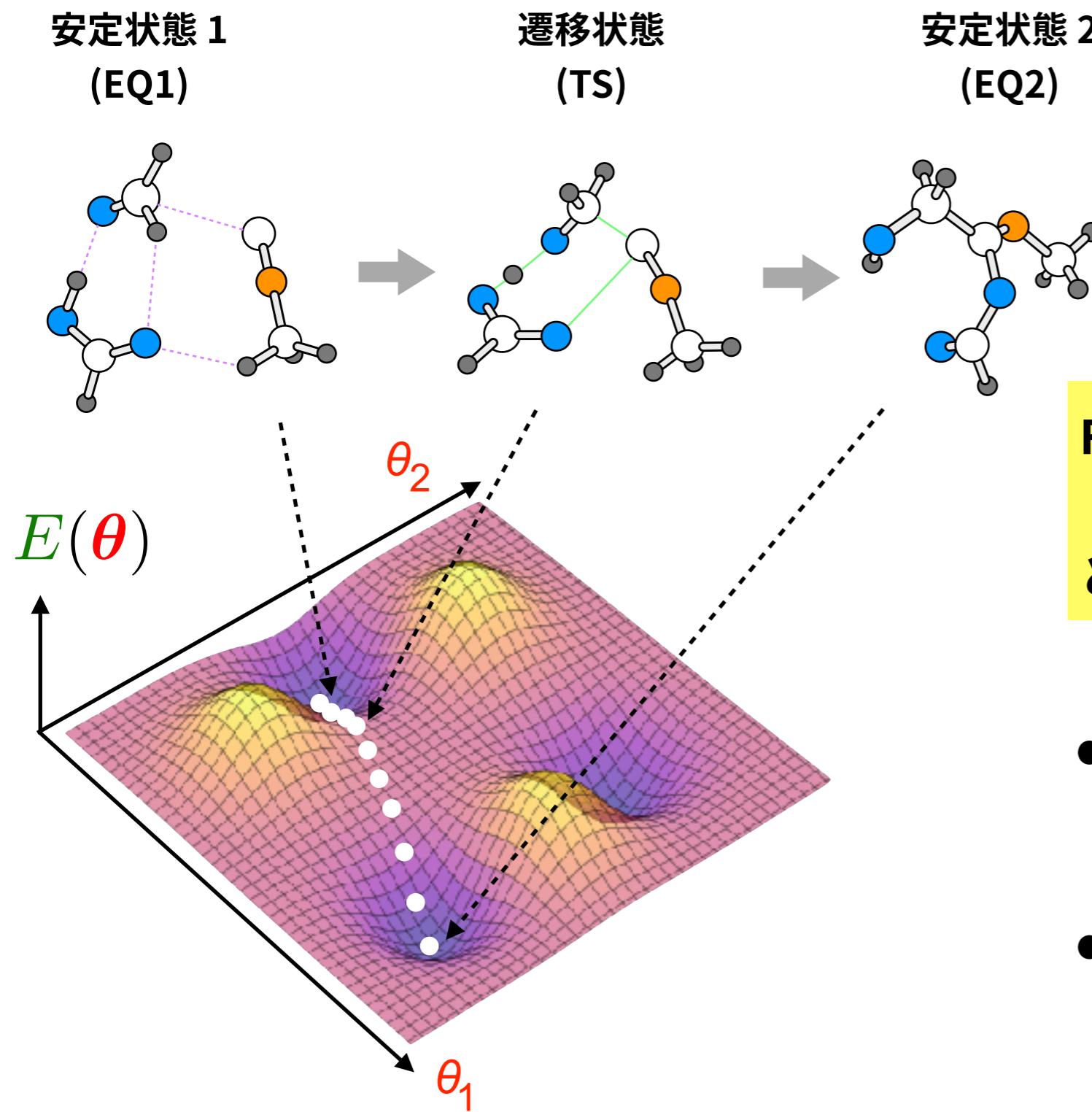


**解くべき問題：**  
**多峰性の多変数関数の  
極小解と鞍点の「全列挙」**

**そんなの無理では！？**

1999年、有名な計算化学の教科書でFrank Jensenは「変数の数が3ないし4を超えると化学反応の遷移状態をすべて求めることは不可能である」と述べた。

# 情報科学から見た反応経路の自動探索



**解くべき問題：**  
多峰性の多変数関数の  
極小解と鞍点の「全列挙」

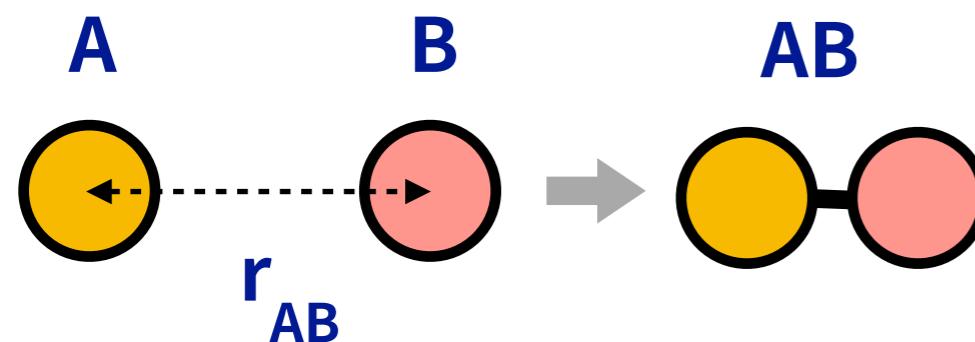
Frank Jensenの教科書は2007年に  
「大野・前田らの方法で可能になった」と書き換えられた！

- 極小解(EQ)および鞍点(TS)を求める → 「AFIR」
- 全列挙・制約付き経路探索 → 「GRRM」

# 極小解(EQ)および鞍点(TS)を求める→「AFIR」

そもそも「PES」の多峰性がなぜ生じるかを考えてみる...

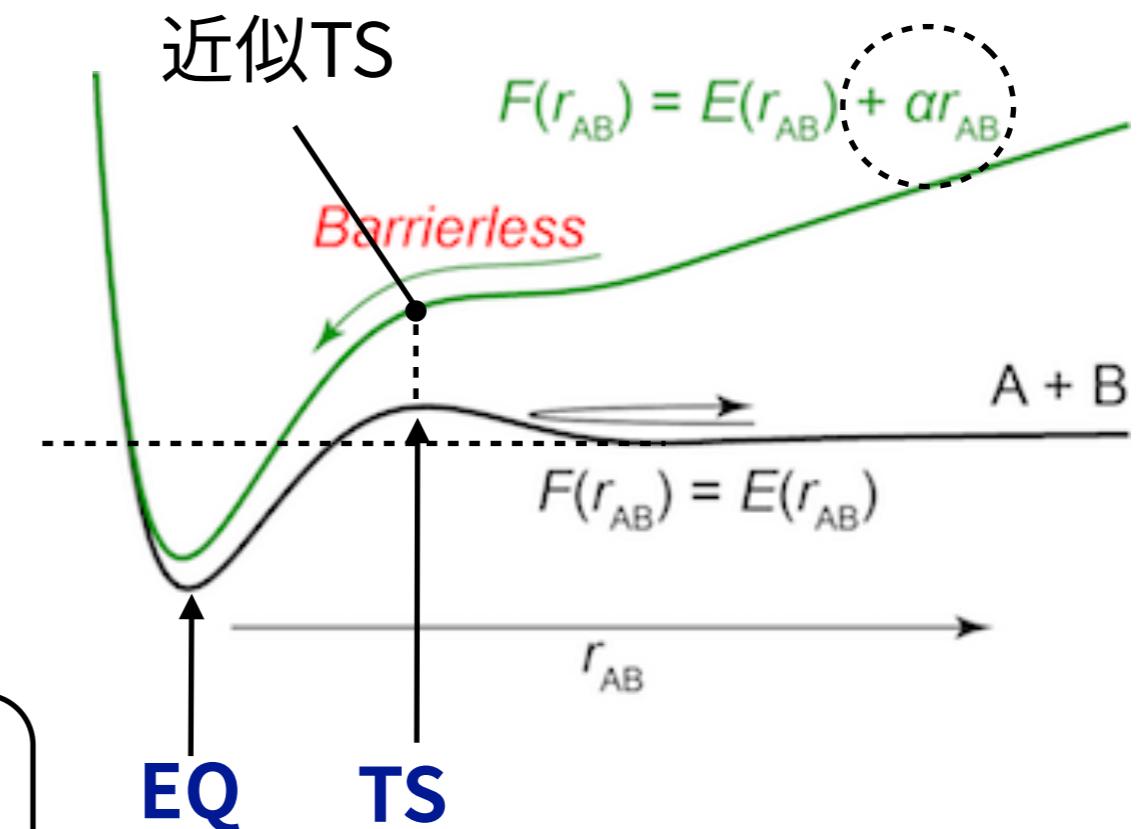
2つの原子AとBが反応して分子ABになるときエネルギー障壁が現れるから



距離に応じたエネルギー(力)を  
人工的に加えることに相当

$$E(r_{AB})$$

$$F(r_{AB}) = E(r_{AB}) + \alpha r_{AB}$$

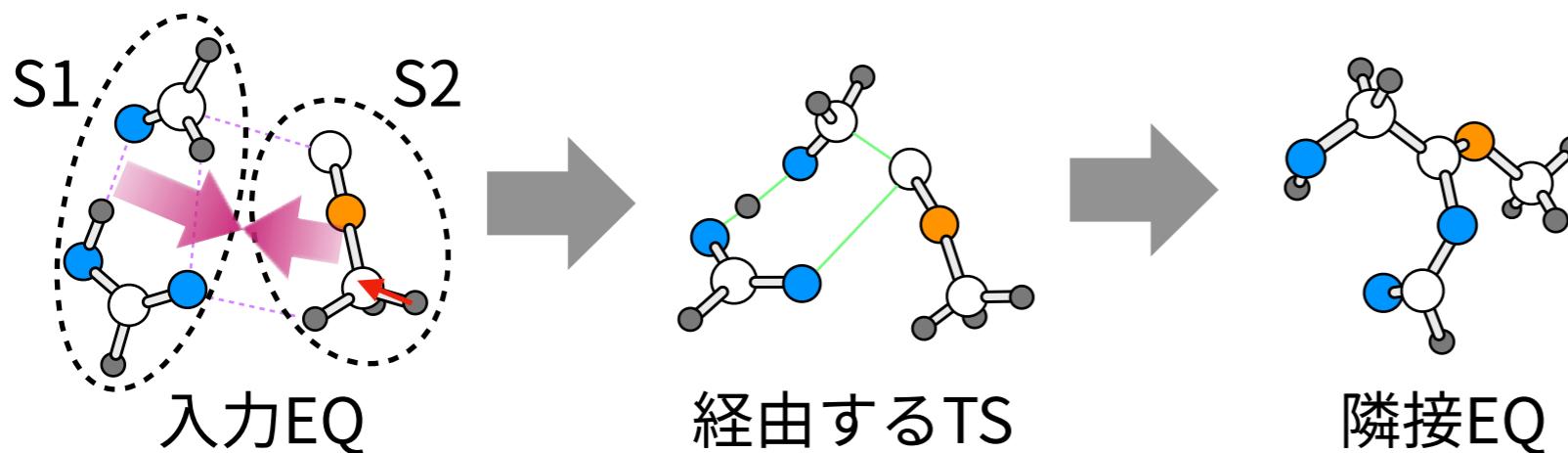


距離に応じたゲタをはかせると  
どこかで単調減少関数になり  
難しいHill Climbingが不要に

# 極小解(EQ)および鞍点(TS)を求める→「AFIR」

## Artificial Force Induced Reaction (人工力誘起反応法)

各原子を原子間の距離が小さくなる方向へ人工力Fで押す(or引く)



- 隣接EQと経由TSが得られる
- Fを系統的に変えて隣接EQを網羅的に探索

原子集合 S1 と S2 の間のAFIR関数

$$F(\theta) = E(\theta) + \alpha \frac{\sum_{i \in S_1} \sum_{j \in S_2} [(R_i + R_j)/r_{ij}]^p \cdot r_{ij}}{\sum_{i \in S_1} \sum_{j \in S_2} [(R_i + R_j)/r_{ij}]^p}$$

力Fの強さ  $\alpha = \frac{\gamma}{\left[ 2^{-1/p} - \left( 1 + \sqrt{1 + \gamma/\varepsilon} \right)^{-1/p} \right] R_0}$

$R_i$  共有結合半径  
 $r_{ij}$  距離

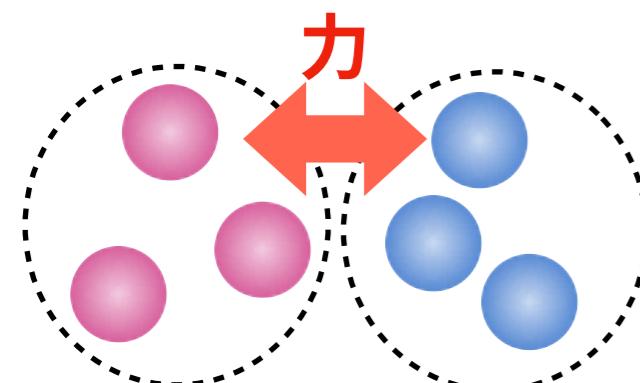
$\gamma$  パラメタ  
この近似的な上界値以下の反応障壁で遷移できる反応経路を網羅的に探索

# 極小解(EQ)および鞍点(TS)を求める→「AFIR」

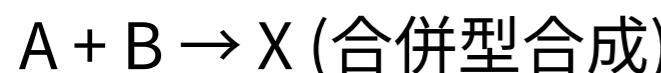
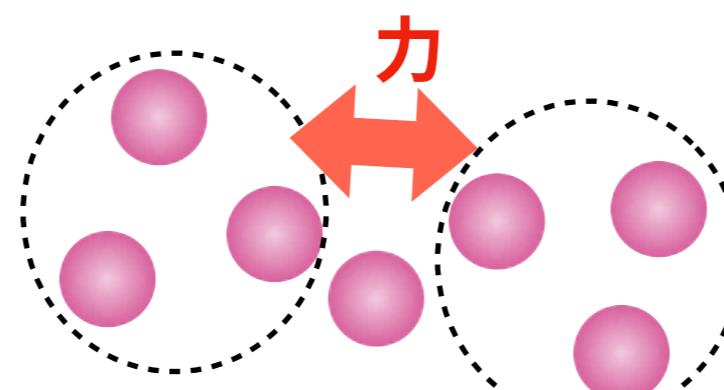
## Artificial Force Induced Reaction (人工力誘起反応法)

各原子を原子間の距離が小さくなる方向へ人工力Fで押す(or引く)

MC-AFIR

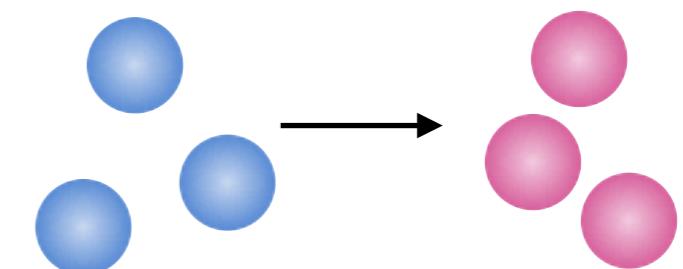


SC-AFIR



多成分の間に力をかける

DS-AFIR



出発物質と産物が  
与えられる二端点経路  
を求める場合に使う

单一成分の内の原子団の  
間に力をかける

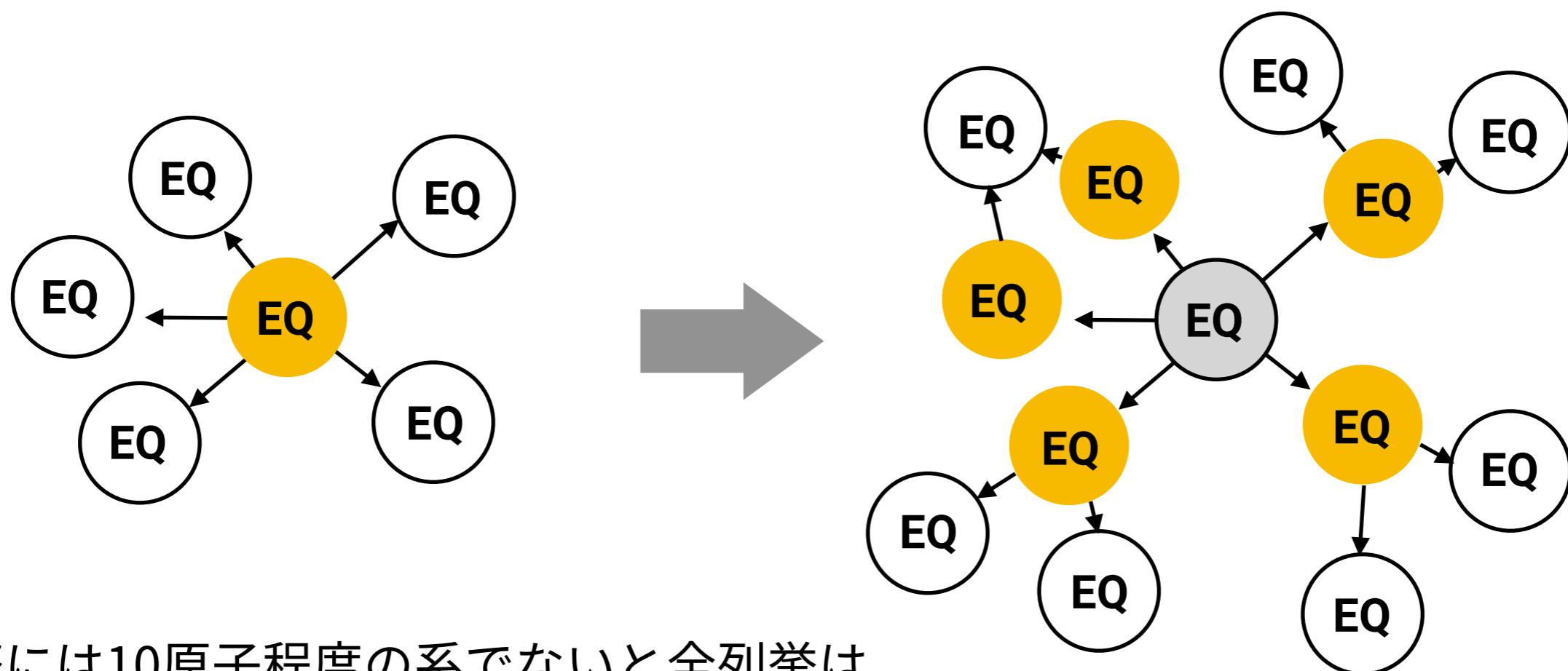
ADDF

SC-AFIRと同じ目的だが  
AFIRと異なる手法

# 全列挙・制約付き経路探索→「GRRM」

## Global Reaction Route Mapping (グローバル反応経路マッピング)

出発状態から遷移可能なEQをAFIRで列挙し、新たなEQが出なくなるまで再帰的にこの処理を適用



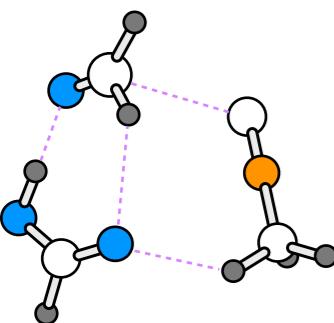
実際には10原子程度の系でないと全列挙は  
数が多くて無理なので適宜必要に応じて制約  
→ GRRM20では速度定数行列収縮法(RCMC)が利用可能

# ICReDD: 化学反応のデザインと発見



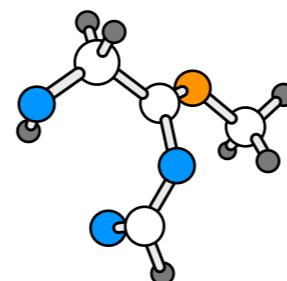
*Chemical Reaction*

EQ1



*How we can have this?*

EQ2

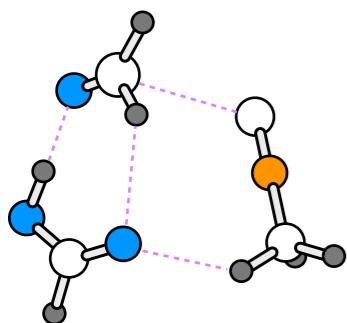


# ICReDD: 化学反応のデザインと発見

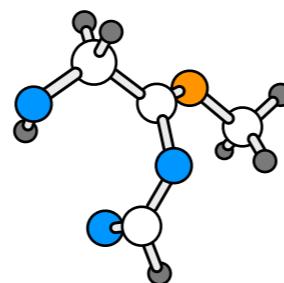


*Chemical Reaction*

EQ1



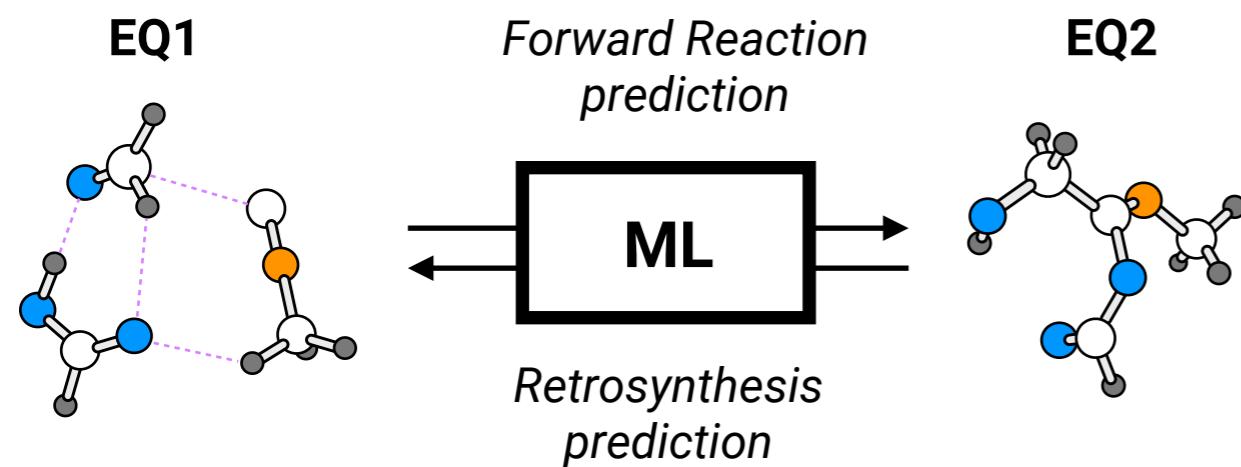
EQ2



# ICReDD: 化学反応のデザインと発見



## Chemical Reaction

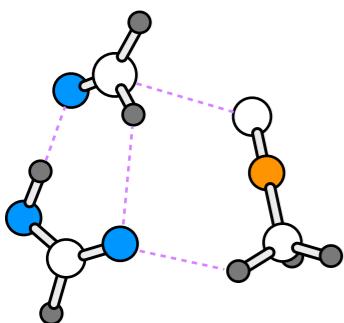


# ICReDD: 化学反応のデザインと発見

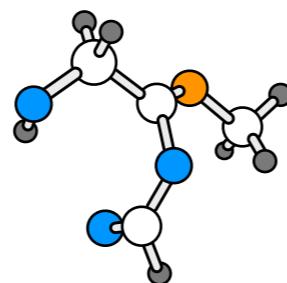


*Chemical Reaction*

EQ1



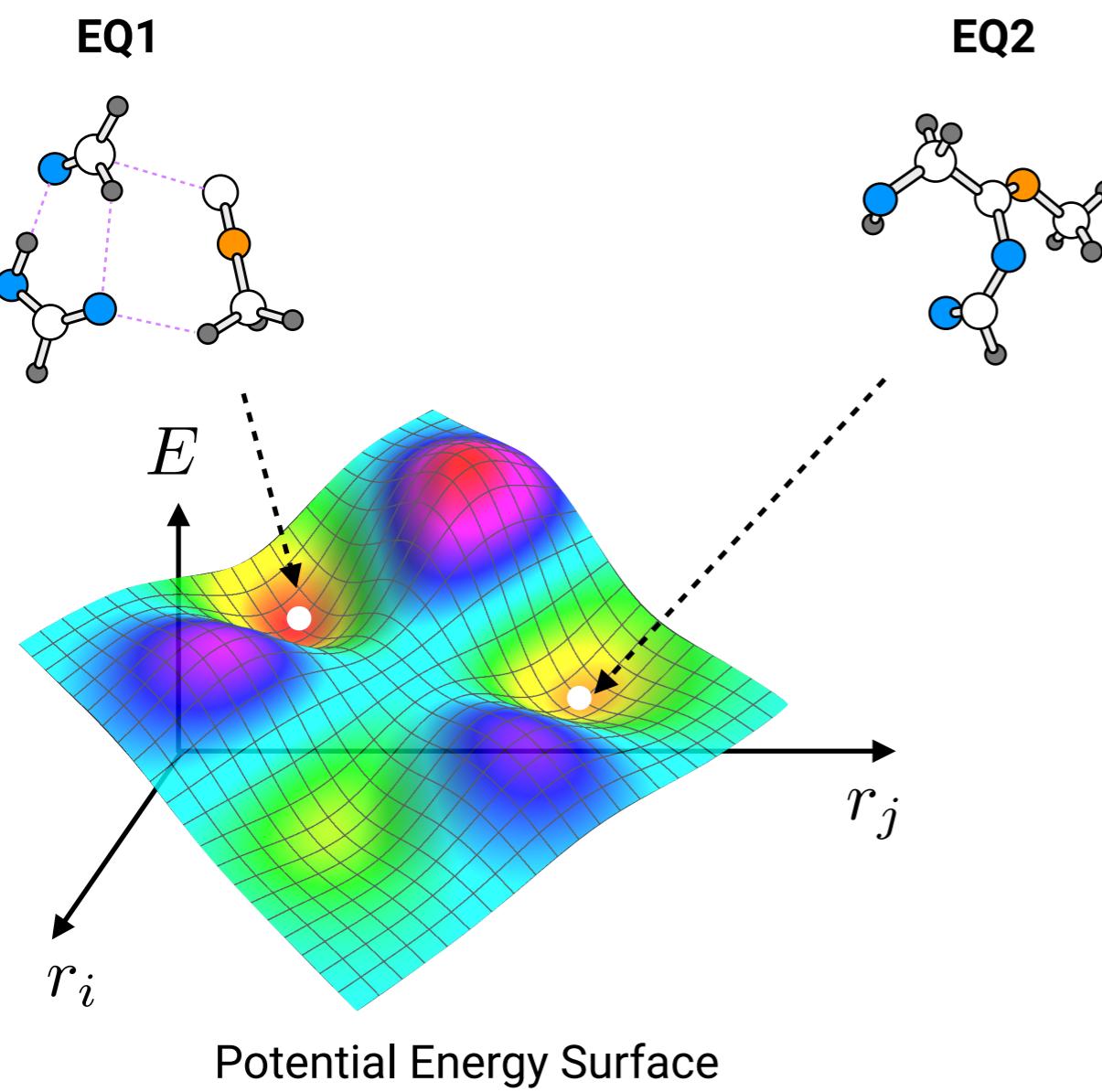
EQ2



# ICReDD: 化学反応のデザインと発見



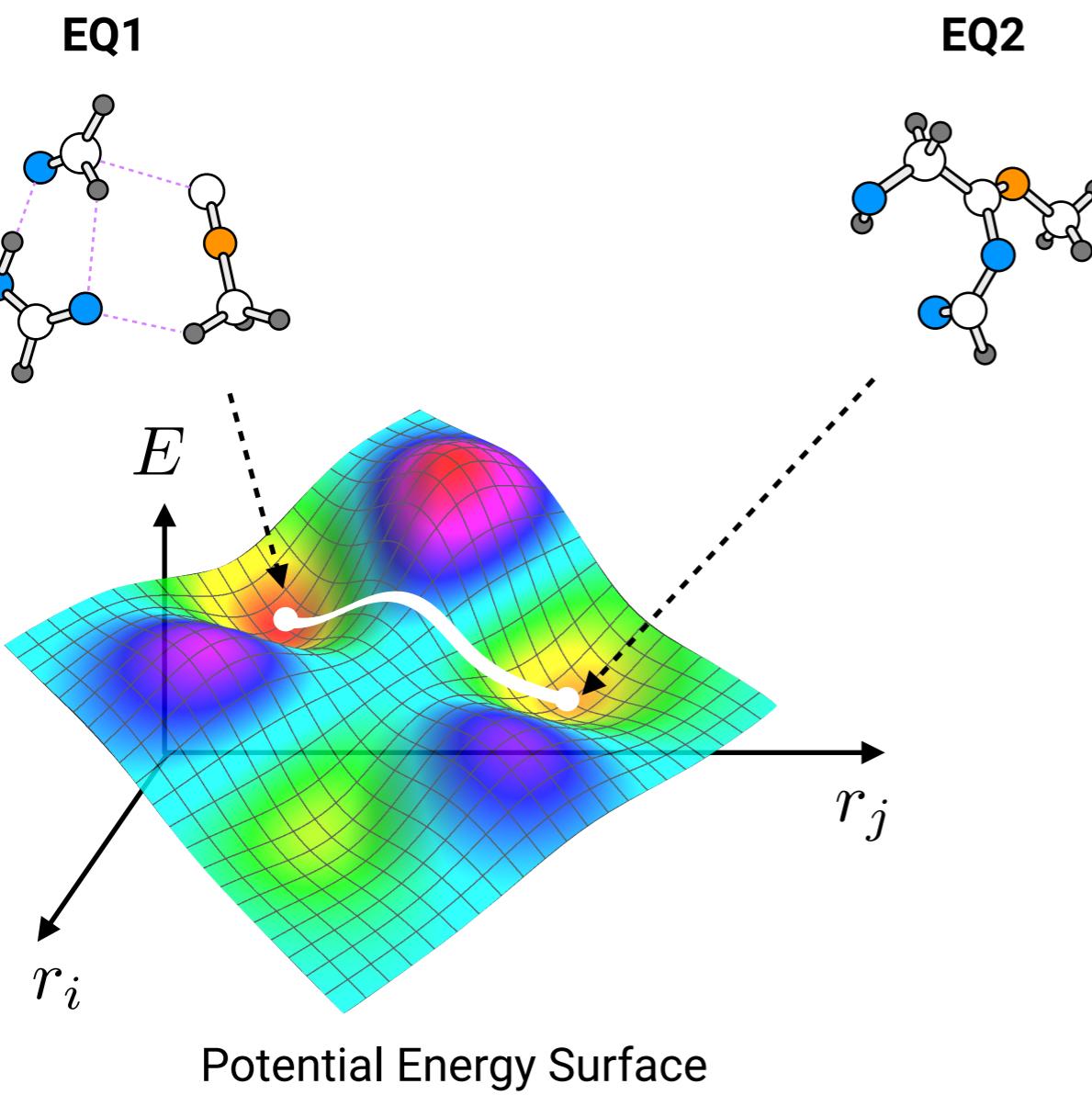
*Chemical Reaction*



# ICReDD: 化学反応のデザインと発見



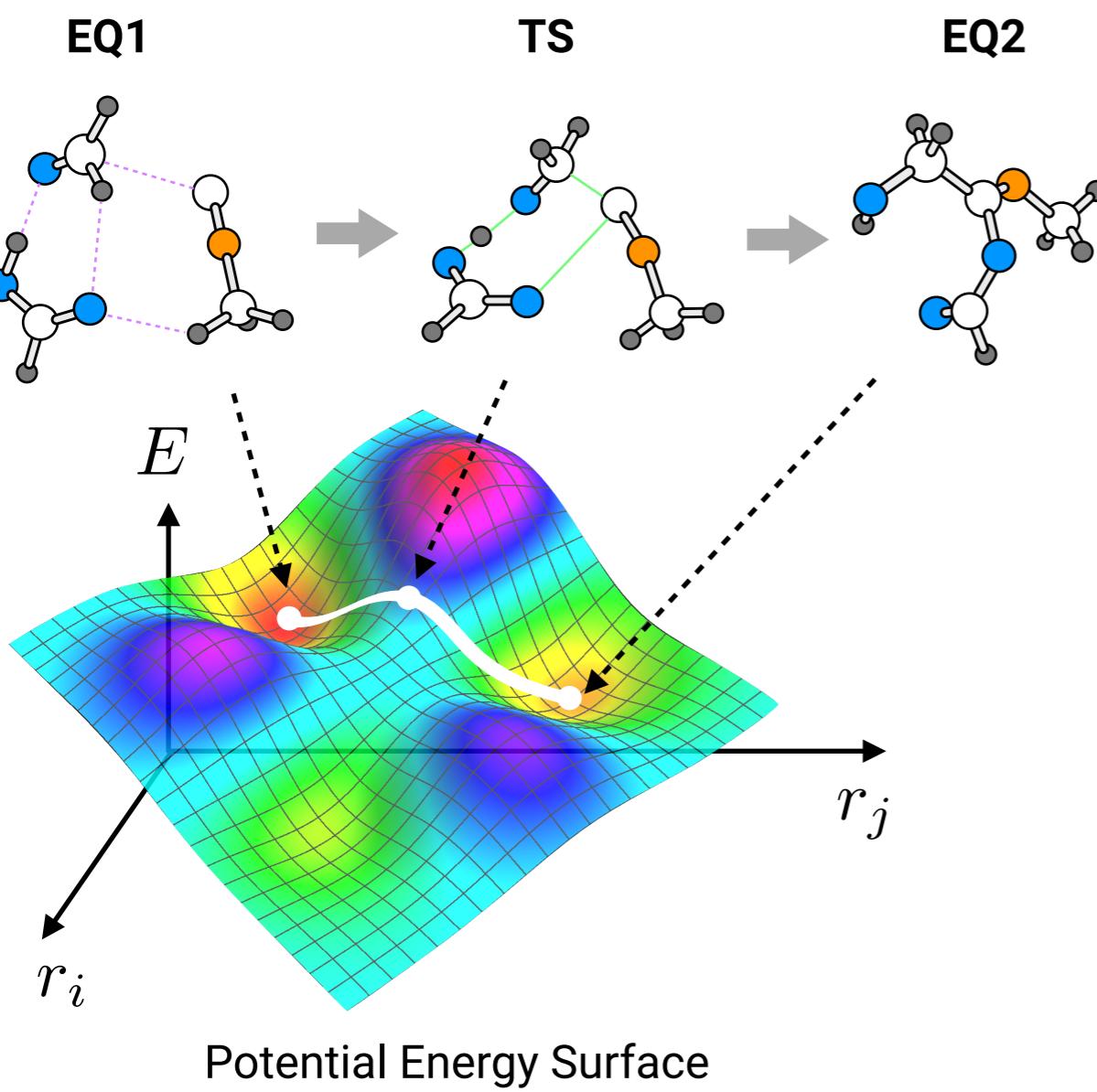
*Chemical Reaction*



# ICReDD: 化学反応のデザインと発見



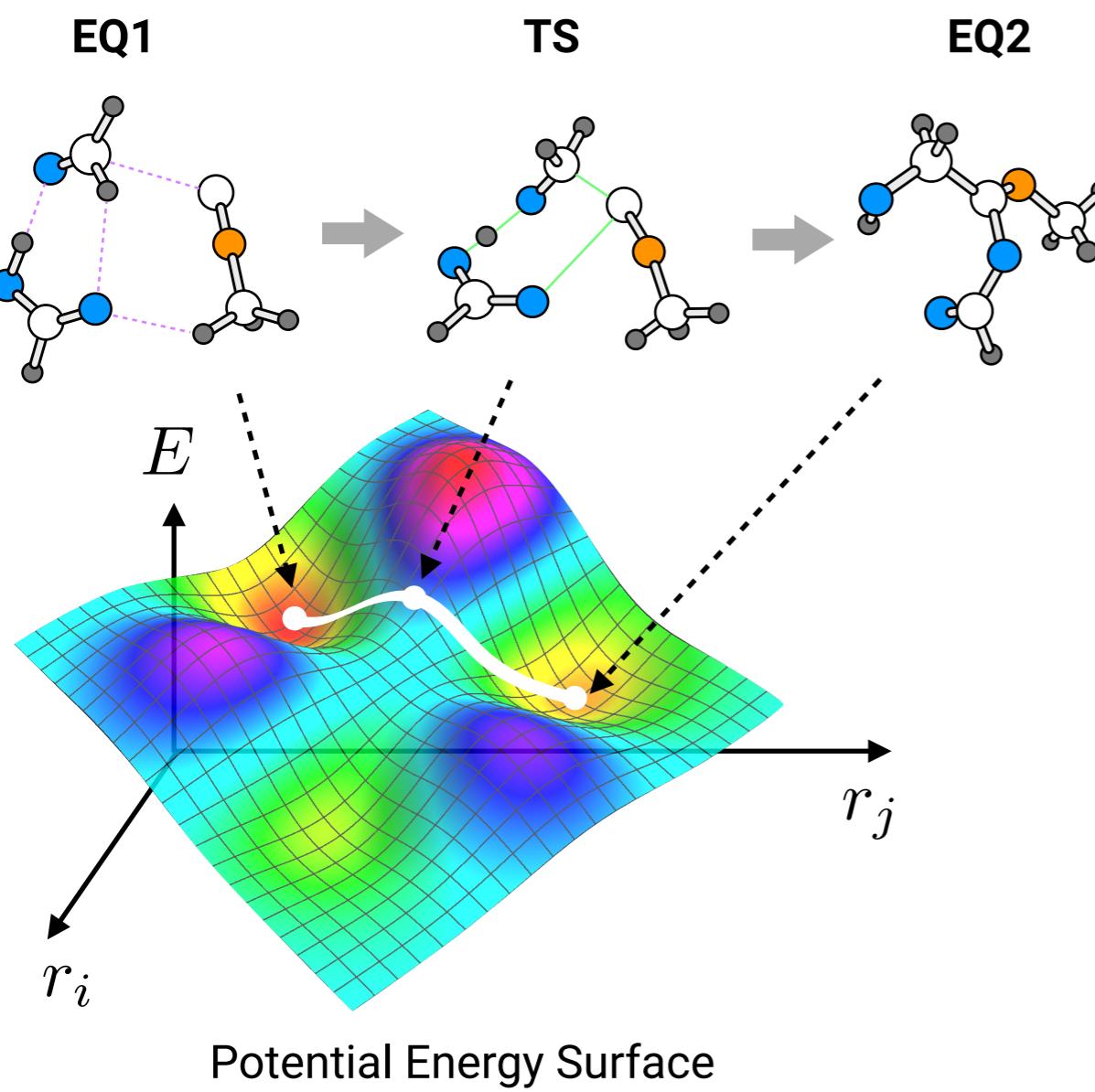
*Chemical Reaction*



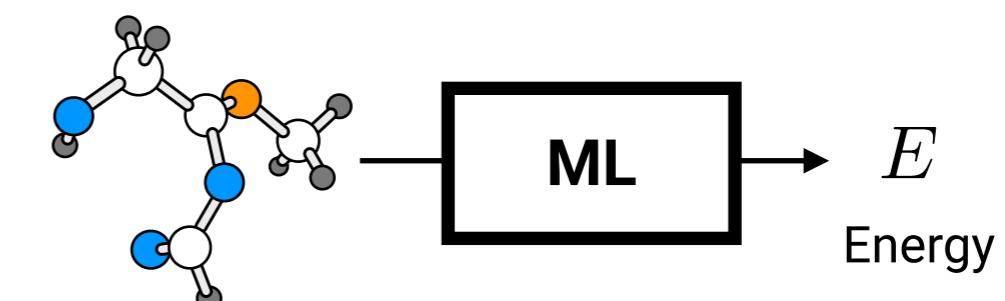
# ICReDD: 化学反応のデザインと発見



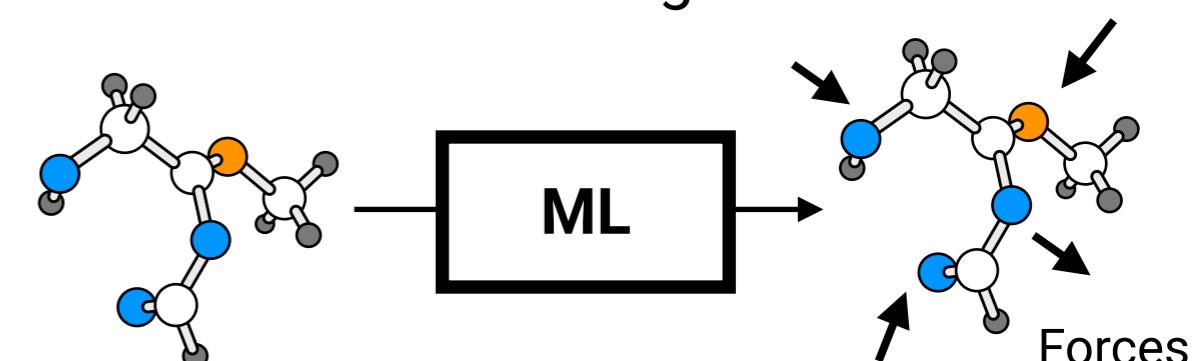
## Chemical Reaction



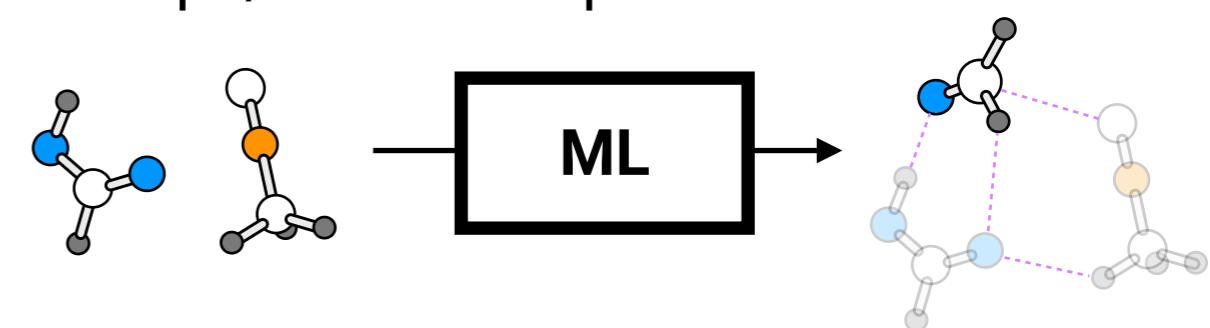
- Fill any gap between theory and experiments (reality) by data?
- Acceleration by ML potential?



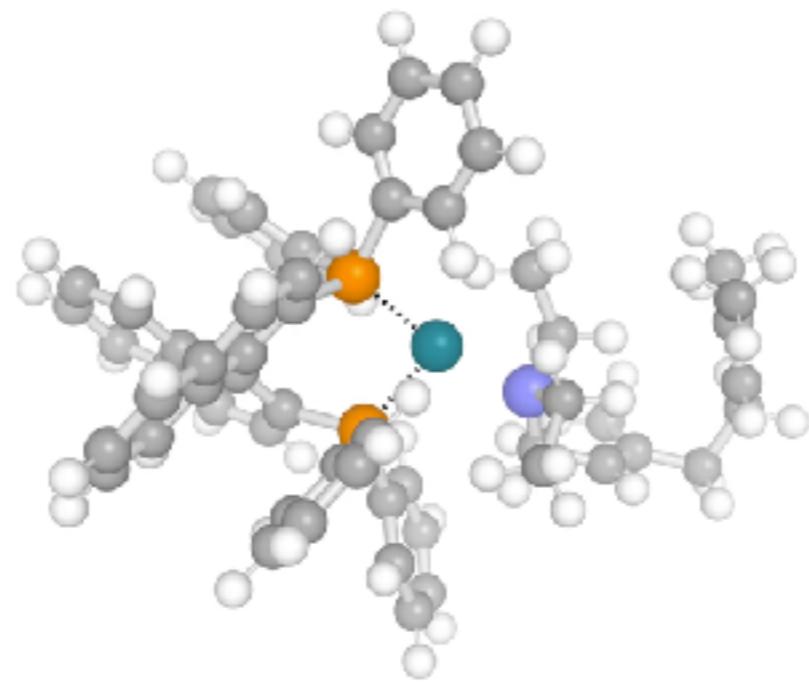
- Artificial force learning?



- Scope/network expansion?



分子の形を見て反応が起こるようにうまく押したり引いたりは学習できる…？



# 量子化学 + 機械学習の融合を目指して

ARTICLE

<https://doi.org/10.1038/s41467-020-19093-1>

OPEN

 Check for updates

## Quantum chemical accuracy from density functional approximations via machine learning

Mihail Bogojeski<sup>1,9</sup>, Leslie Vogt-Maranto<sup>1,9</sup>, Mark E. Tuckerman<sup>2,3,4</sup>, Klaus-Robert Müller<sup>1,5,6</sup> & Kieron Burke<sup>1,7,8</sup>

# CHEMICAL REVIEWS

[pubs.acs.org/CR](https://pubs.acs.org/CR)



Review

## Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems

John A. Keith,\* Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller,\* and Alexandre Tkatchenko\*



Cite This: <https://doi.org/10.1021/acs.chemrev.1c00107>



Read Online

# もしくは、シミュレーション+機械学習の融合

*Annu. Rev. Phys. Chem.* 71:361–90 (2020)



*Annual Review of Physical Chemistry*  
Machine Learning for  
Molecular Simulation

Frank Noé,<sup>1,2,3</sup> Alexandre Tkatchenko,<sup>4</sup>  
Klaus-Robert Müller,<sup>5,6,7</sup> and Cecilia Clementi<sup>1,3,8</sup>

PNAS (2020)

*Nat. Rev. Chem.* 4: 347–358 (2020)

NATURE REVIEWS | CHEMISTRY

PERSPECTIVES

Exploring chemical compound  
space with quantum-based machine  
learning

O. Anatole von Lilienfeld, Klaus-Robert Müller and Alexandre Tkatchenko

**Abstract** | Rational design of compounds with specific properties requires understanding and fast evaluation of molecular properties throughout chemical compound space — the huge set of all potentially stable molecules. Recent advances in combining quantum-mechanical calculations with machine learning

## The frontier of simulation-based inference

Kyle Cranmer<sup>a,b,1</sup> , Johann Brehmer<sup>a,b</sup> , and Gilles Louppe<sup>c</sup>

<sup>a</sup>Center for Cosmology and Particle Physics, New York University, New York, NY 10003; <sup>b</sup>Center for Data Science, New York University, New York, NY 10011;  
and <sup>c</sup>Montefiore Institute, University of Liège, B-4000 Liège, Belgium

Edited by Jitendra Malik, University of California, Berkeley, CA, and approved April 10, 2020 (received for review November 4, 2019)

**Many domains of science have developed complex simulations to describe phenomena of interest. While these simulations provide high-fidelity models, they are poorly suited for inference and lead to challenging inverse problems. We review the rapidly developing field of simulation-based inference and identify the forces giving additional momentum to the field. Finally, we describe how the frontier is expanding so that a broad audience can appreciate the profound influence these developments may have on science.**

the simulator—is being recognized as a key idea to improve the sample efficiency of various inference methods. A third direction of research has stopped treating the simulator as a black box and focused on integrations that allow the inference engine to tap into the internal details of the simulator directly.

Amidst this ongoing revolution, the landscape of simulation-based inference is changing rapidly. In this review we aim to provide the reader with a high-level overview of the basic ideas

# 手続き的・記号的操作の機械学習技術も飛躍的に発展

## Neural Abstract Machines & Program Induction

<https://uclnlp.github.io/nampi/>

*Machine intelligence capable of learning complex procedural behavior, inducing (latent) programs, and reasoning with these programs is a key to solving artificial intelligence. Recently, there have been a lot of success stories in the deep learning community related to learning neural networks capable of using trainable memory abstractions.*

- Differentiable Neural Computers / Neural Turing Machines (Graves+ 2014)
- Memory Networks (Weston+ 2014)
- Pointer Networks (Vinyals+ 2015)
- Neural Stacks (Grefenstette+ 2015, Joulin+ 2015)
- Hierarchical Attentive Memory (Andrychowicz+ 2016)
- Neural Program Interpreters (Reed+ 2016)
- Neural Programmer (Neelakantan+ 2016)
- DeepCoder (Balog+ 2016)
- :

手続き的・記号的操作も学習できるプログラムとして扱えるようになってきた

Computer-Aided Synthetic Planning

International Edition: DOI: 10.1002/anie.201506101  
German Edition: DOI: 10.1002/ange.201506101

### Computer-Assisted Synthetic Planning: The End of the Beginning

Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A. Grzybowski\*

Angew. Chem. Int. Ed. 2016, 55, 5904–5937



AI-Assisted Synthesis Very Important Paper

Reaxys®

SCI-FINDER®  
A CAS SOLUTION

CHEMATICAA

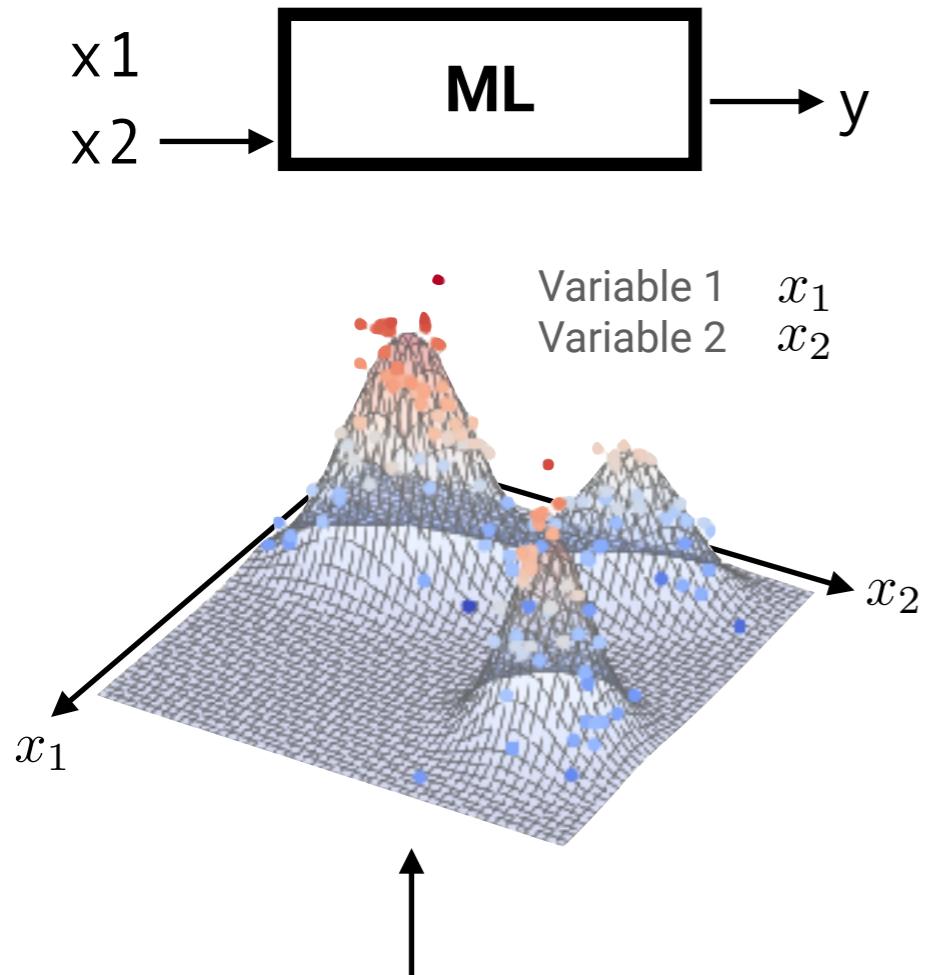
International Edition: DOI: 10.1002/anie.201912083  
German Edition: DOI: 10.1002/ange.201912083

### Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning

Tomasz Badowski, Ewa P. Gajewska, Karol Molga, and Bartosz A. Grzybowski\*

Angew. Chem. Int. Ed. 2019, 58, 1–7

# 帰納バイアスのデザイン：明示的知識+機械学習の融合



現代的MLでは  
入力変数の増大(キッチンシンク化)  
とパラメタモデル数の肥大化により  
モデルの自由度が高すぎる…

汎用のMLモデルではフィッティングに使う  
モデルの自由度が高すぎて  
Underspecificationのリスクが高すぎる  
(現象を理解したい自然科学分野では大問題)



理論化学など既知の知識を総動員して「無意  
味なフィッティング結果に陥らないように」  
モデルの自由度を予め上手に制約しておく

- ✓ 機械学習モデルの成功のためには多かれ少なかれ  
モデルの帰納バイアスが目前の問題に適合してい  
る必要がある (既知なことは学習する必要がない)
- ✓ 汎用性の高いモデルを高い確度で学習するには  
相応の膨大なデータが必要 (多くの場合、難しい)

# グレイボックス最適化：論理推論と統計的予測の融合

## Theory-driven 【演繹・合理論】

- 対象現象の複雑化
- シミュレーション技法も複雑化
- "経験的に決める"パラメタや初期値
- 汎関数、交換相関項の設計

## (人工知能分野)

- 知識ベースと論理推論(記号AI)の限界
- 厳密推論や探索の計算爆発(NP困難性)
- 大量データの知識化の問題
- 制約プログラミングや組合せ最適化

→ データ同化、模倣学習、論理合成、etc

→ 表現学習、モデルベース最適化・強化  
学習、メタ学習、生成モデル、etc

→ 新たな方法論へ？

## Data-driven 【帰納・経験論】

- 小サンプル・低カウントの問題
- 帰納バイアスのモデルエンコード
- 外挿の低信頼性と探索
- Blackbox性・解釈性の問題

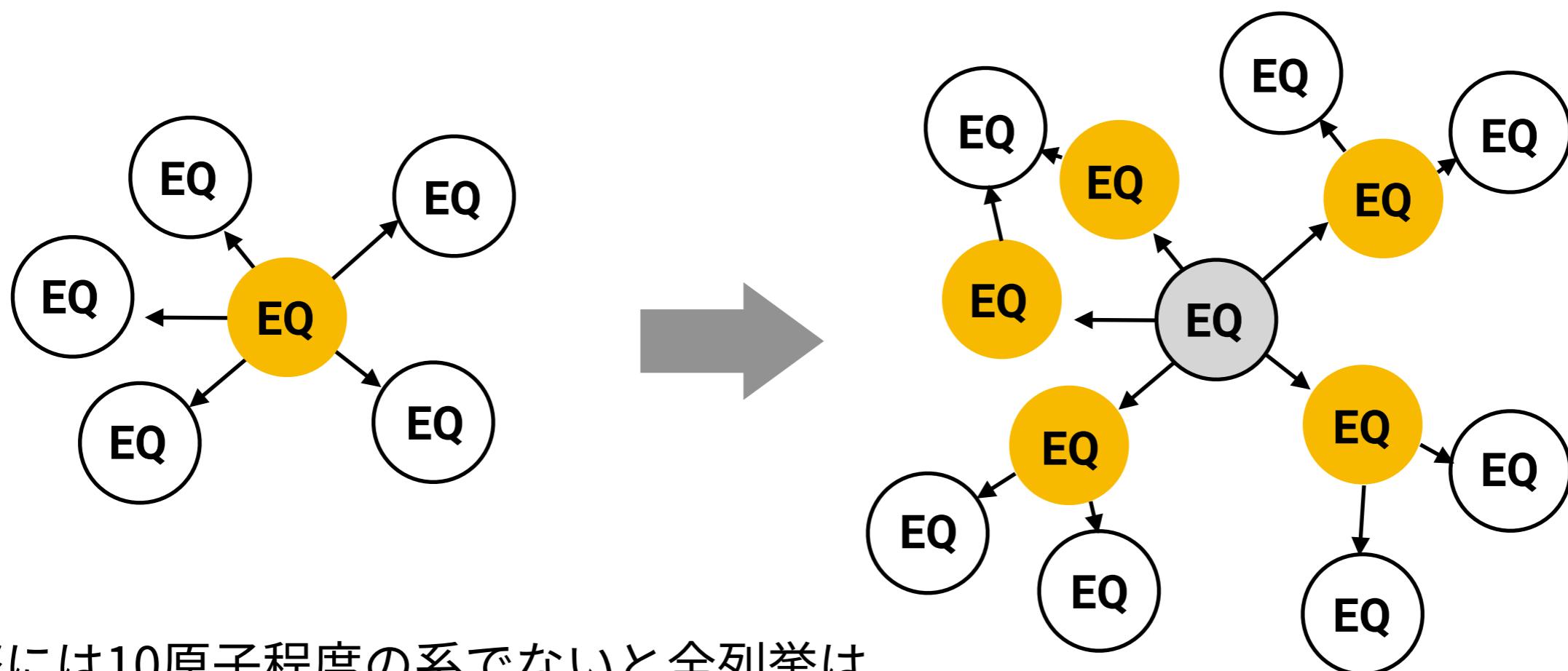
## (人工知能分野)

- Data-Driven手法(機械学習)と人間の論理的思考との大きなギャップ
- Dataがない領域の探索や「ひらめき」
- モデル適用範囲と信頼性・安全性

# 全列挙・制約付き経路探索→「GRRM」

## Global Reaction Route Mapping (グローバル反応経路マッピング)

出発状態から遷移可能なEQをAFIRで列挙し、新たなEQが出なくなるまで再帰的にこの処理を適用



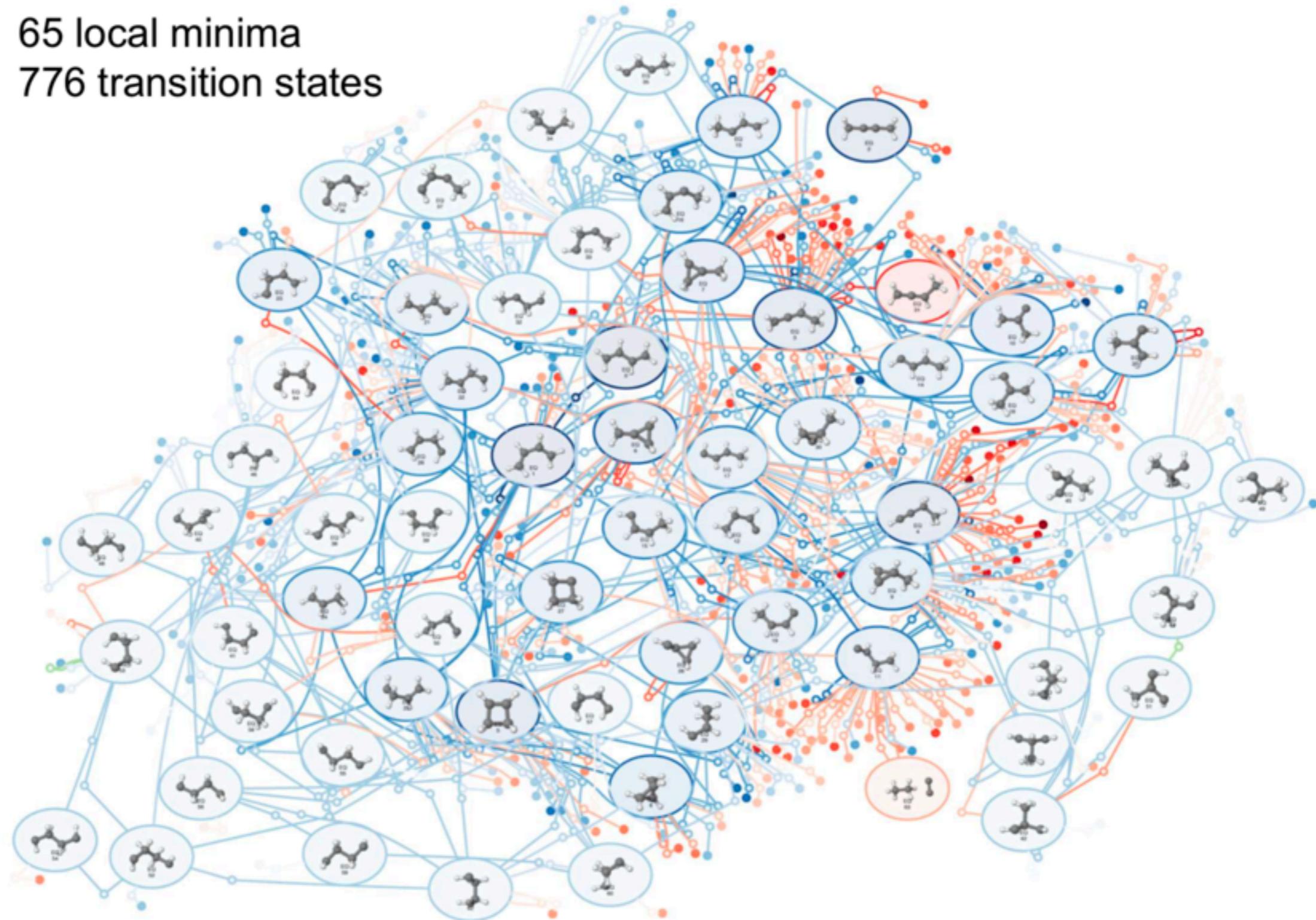
実際には10原子程度の系でないと全列挙は  
数が多くて無理なので適宜必要に応じて制約  
→ GRRM20では速度定数行列収縮法(RCMC)が利用可能

# "反応経路ネットワーク"

## Global reaction route map of C<sub>4</sub>H<sub>6</sub>

65 local minima

776 transition states



# 反応経路ネットワークはBioinfoで散々研究してきた対象



Biochemical Pathways Roche.com Contact Share [✉](#) [f](#) [t](#) [in](#) [g+](#)

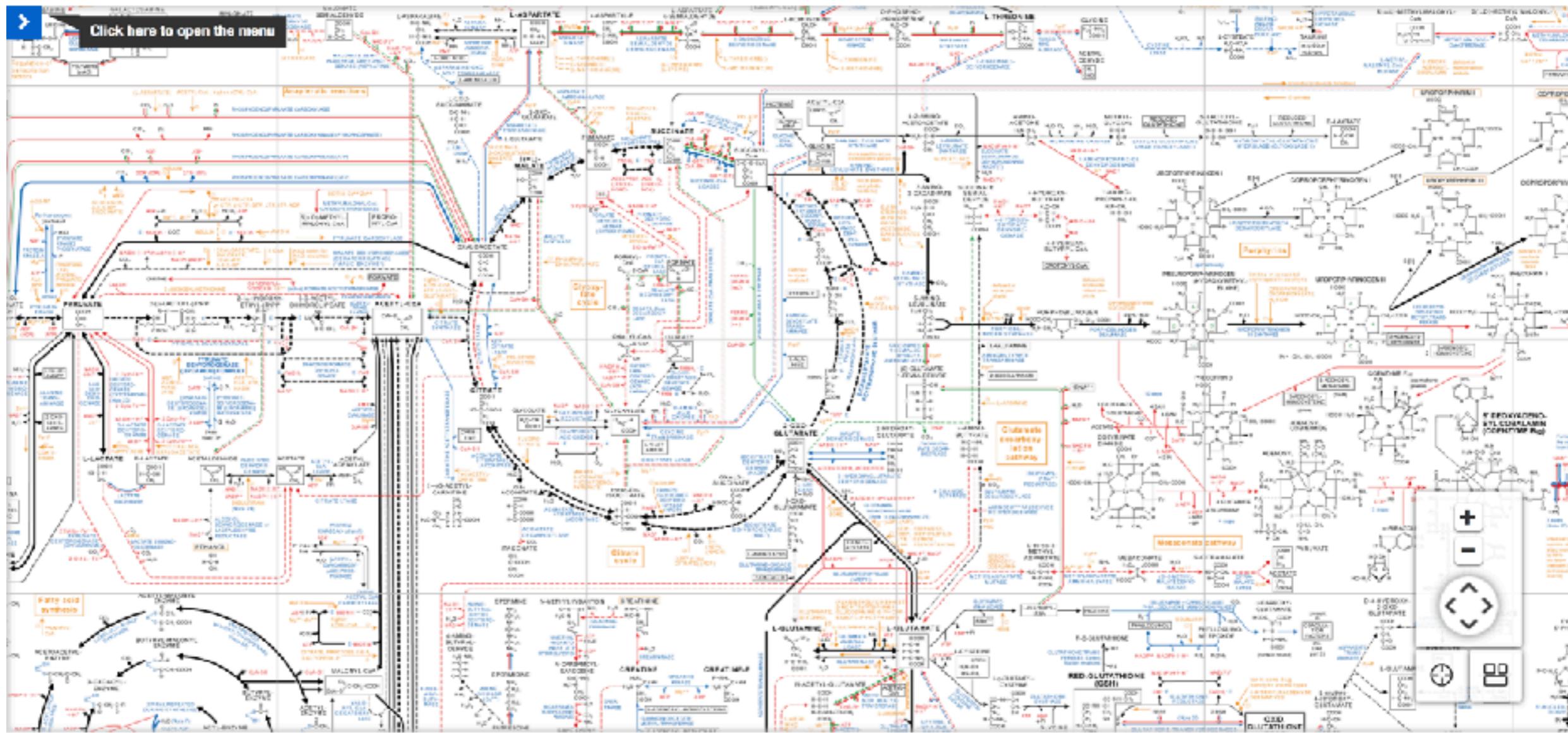


## Pathway Commons

Access and discover data integrated from public pathway and interactions databases.



Part 1: Metabolic Pathways Part 2: Cellular and Molecular Processes



# 化学のホットトピックであるだけでなく…

Science

**REVIEW**

## Inverse molecular design using machine learning: Generative models for matter engineering

Benjamin Sanchez-Lengeling<sup>1</sup> and Alán Aspuru-Guzik<sup>2,3,4\*</sup>

## Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

Andrew F. Zahrt<sup>5</sup>, Jeremy J. Henle<sup>6</sup>, Brennan T. Rose, Yang Wang,  
William T. Darrow, Scott E. Denmark<sup>†</sup>

nature reviews chemistry

**REVIEWS**

## Synthetic organic chemistry driven by artificial intelligence

A. Filipa de Almeida<sup>1</sup>, Rui Moreira<sup>1</sup> and Tiago Rodrigues<sup>1,2\*</sup>

**PERSPECTIVES**

## Exploring chemical compound space with quantum-based machine learning

O. Anatole von Lilienfeld, Klaus-Robert Müller and Alexandre Tkatchenko<sup>3</sup>

nature

**REVIEW**

## Machine learning for molecular and materials science

Keith T. Badger<sup>1</sup>, Daniel W. Davies<sup>2</sup>, Hugh Cartwright<sup>3</sup>, Olexandr Isayev<sup>4,\*</sup> & Aron Walsh<sup>5,6\*</sup>

## Planning chemical syntheses with deep neural networks and symbolic AI

Marwin H. S. Segler<sup>1,3</sup>, Mike Preuss<sup>3</sup> & Mark P. Waller<sup>4</sup>

## Holistic prediction of enantioselectivity in asymmetric catalysis

Iolene P. Reid<sup>3</sup> & Matthew S. Sigman<sup>1\*</sup>

## Bayesian reaction optimization as a tool for chemical synthesis

Benjamin J. Shields<sup>1</sup>, Jason Stevens<sup>2</sup>, Jun Li<sup>2</sup>, Marvin Parashram<sup>1</sup>, Farhan Damani<sup>3</sup>, Jesus I. Martinez Alvarado<sup>1</sup>, Jacob M. Janey<sup>2</sup>, Ryan P. Adams<sup>2,3</sup> & Abigail G. Doyle<sup>1,2\*</sup>

# 化学のホットトピックであるだけでなく…

**Computer Chemistry**

How to cite: *Angew. Chem. Int. Ed.* **2020**, *59*, 18860–18865  
 International Edition: doi.org/10.1002/anie.202008366  
 German Edition: doi.org/10.1002/ange.202008366

## Molecular Machine Learning: The Future of Synthetic Chemistry?

Philipp M. Pflüger and Frank Glorius\*

**Reaction Prediction**

International Edition: DOI: 10.1002/anie.201803562  
 German Edition: DOI: 10.1002/ange.201803562

## Machine Learning for Organic Synthesis: Are Robots Replacing Chemists?

Boris Maryasin, Philipp Marquetand, and Nuno Maulide\*



**AI-Assisted Synthesis Very Important Paper**

International Edition: DOI: 10.1002/anie.201912083  
 German Edition: DOI: 10.1002/ange.201912083

## Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning

Tomasz Badowski, Ewa P. Gajewska, Karol Molga, and Bartosz A. Grzybowski\*

**Computer-Aided Synthetic Planning**

International Edition: DOI: 10.1002/anie.201506101  
 German Edition: DOI: 10.1002/ange.201506101

## Computer-Assisted Synthetic Planning: The End of the Beginning

Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A. Grzybowski\*



# 機械学習分野のホットトピックでもある！

## NeurIPS 2020

- *Self-Supervised Graph Transformer on Large-Scale Molecular Data*
- *RetroXpert: Decompose Retrosynthesis Prediction Like A Chemist*
- *Reinforced Molecular Optimization with Neighborhood-Controlled Grammars*
- *Autofocused Oracles for Model-based Design*
- *Barking Up the Right Tree: an Approach to Search over Molecule Synthesis DAGs*
- *On the Equivalence of Molecular Graph Convolution and Molecular Wave Function with Poor Basis Set*
- *CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models*

## ICLR 2020, 2021

- *Directional Message Passing for Molecular Graphs*
- *GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation*
- *Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space*
- *A Fair Comparison of Graph Neural Networks for Graph Classification*
- *MARS: Markov Molecular Sampling for Multi-objective Drug Discovery*
- *Practical Massively Parallel Monte-Carlo Tree Search Applied to Molecular Design*
- *Learning Neural Generative Dynamics for Molecular Conformation Generation*
- *Conformation-Guided Molecular Representation with Hamiltonian Neural Networks*
- *Symmetry-Aware Actor-Critic for 3D Molecular Design*

## ICML 2020, 2021

- *A Graph to Graphs Framework for Retrosynthesis Prediction*
- *Hierarchical Generation of Molecular Graphs using Structural Motifs*
- *Learning to Navigate in Synthetically Accessible Chemical Space Using Reinforcement Learning*
- *Reinforcement Learning for Molecular Design Guided by Quantum Mechanics*
- *Multi-Objective Molecule Generation using Interpretable Substructures*
- *Improving Molecular Design by Stochastic Iterative Target Augmentation*
- *A Generative Model for Molecular Distance Geometry*
- *GraphDF: A Discrete Flow Model for Molecular Graph Generation*
- *An End-to-End Framework for Molecular Conformation Generation via Bilevel Programming*
- *Equivariant message passing for the prediction of tensorial properties and molecular spectra*
- *Learning Gradient Fields for Molecular Conformation Generation*
- *Self-Improved Retrosynthetic Planning*

# 今日のテーマ

- **自己紹介 (機械学習と自然科学の境界)**
- **機械学習とは新しいプログラミングの方法**
- **機械学習屋は一体何が楽しいのか？**
  - 分子の表現と機械学習
  - グレイボックス最適化 (演繹 + 帰納)：論理学と統計学の融合？
- **自然科学研究で機械学習を使おうとすると必ずぶつかる本当に難しい問題**
  - データモデリングと予測アルゴリズム (The Two Cultures)
  - 予測か理解か：Rashomon効果, Underspecification, 解釈多様性
  - 人間の認知バイアスに由来する問題：仮説、失敗、成功バイアス、etc.
- **機械学習から機械発見へ**
  - 「発見」「理解」の道筋は合理化できるのか？自動化できるのか？

# ダークサイドへようこそ

- ✓ これまでの話は主に「量子化学計算によるデータ」でクリーンな世界！  
観測ノイズがない・出力を得るのに必要十分な入力情報が分かっている・いろいろなオープンデータが利用できる・etc.
- ✓ 現実はつらい…
  - 観測ノイズがあり物理的複製が必要(二度測ると値が異なる方が普通)
  - 理論計算に取り入れられてない無数の交絡因子や外乱因子の影響
  - 複雑系では入力変数に何を入れるべきなのかが不明というジレンマ  
→ 入出力関係の機序が分からないから機械学習を使いたいのに必要な情報を入力に入れないと機械学習には擬似相関しか見えない
  - そもそも計測・制御できないたくさんのバックグラウンド因子がある
  - 人間が実験を計画すると得られるデータは常にバイアスを含む  
→ 「良い品質の」必要十分な見本例を作るのは本当に難しい！！

# ダークサイドからのTake Homeメッセージ

**自然科学分野での利活用はMLの技術研磨だけでは成功しない。  
分野専門家との協働が必要不可欠**

# ダークサイドからのTake Homeメッセージ

自然科学分野での利活用はMLの技術研磨だけでは成功しない。  
分野専門家との協働が必要不可欠

- MLがどういう技術なのか**MLの特性と限界**を正しく把握する

# ダークサイドからのTake Homeメッセージ

自然科学分野での利活用はMLの技術研磨だけでは成功しない。  
分野専門家との協働が必要不可欠

- MLがどういう技術なのか**MLの特性と限界**を正しく把握する
- 「データの収集計画(実験計画)と品質保証、適用範囲の理解」  
が“**“data-driven”**の心臓であることをいつも心に

# ダークサイドからのTake Homeメッセージ

自然科学分野での利活用はMLの技術研磨だけでは成功しない。  
分野専門家との協働が必要不可欠

- MLがどういう技術なのか**MLの特性と限界**を正しく把握する
- 「データの収集計画(実験計画)と品質保証、適用範囲の理解」  
が“**“data-driven”**の心臓であることをいつも心に
- 「探索」が目的なら**MLの果たす役割はあくまで一部**と心得る
  - 👍 専門家との協働、分野の専門知識に照らした検証・解釈
  - 👍 シミュレーション・実験自動化・論理推論との融合

# The Two Cultures

<https://projecteuclid.org/euclid.ss/1009213726> (Open Access)

*Statistical Science*  
2001, Vol. 16, No. 3, 199–231

## Statistical Modeling: The Two Cultures

**Leo Breiman**

*Abstract.* There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

# Leo Breiman (1928-2005)が好き！

- UC Berkeley名誉教授
- 2005 SIGKDD Innovation Award
- Probability Theorist → Consultant → Statistician

数学のPhD(UC Berkeley) → 数学科tenure教員(UCLA) → 統計コンサルタント(18年) → 統計学科教授(UC Berkeley)

[https://en.wikipedia.org/wiki/File:Leo\\_Breiman.jpg](https://en.wikipedia.org/wiki/File:Leo_Breiman.jpg)



- CART (Classification and Regression Trees), PIMPLE
- Random Forest
- Arcing (aka Boosting)
- Bagging, Pasting
- ACE (Alternative Conditional Expectations)
- Stacked Generalization (aka Stacking/Blending)
- Nonnegative Garrote (LASSOの前身 for Subset回帰)
- Instability / Stabilization in Model Selection
- Shannon-McMillan-Breiman Theorem (漸近等分割性)
- Kelly-Breiman Strategy (最適な定比例戦略)

If statistics is an applied field and not a minor branch of mathematics,  
then 99% of the published papers are useless exercises.

("Reflections after refereeing papers for NIPS", The Mathematics of Generalization, Ed. D.H. Wolpert, 1995)

# Leo Breiman (1928-2005)が好き！

## THE ANNALS of APPLIED STATISTICS

*AN OFFICIAL JOURNAL OF THE  
INSTITUTE OF MATHEMATICAL STATISTICS*

**Annals of Applied Statistics (Vol. 4,  
No. 4, December 2010)** にBreimanの  
追悼特集があり、色々な関係者が思い出  
や歴史を語っています！  
(もちろんFriedman, Olshen, Stoneも)

**ファン必見！**

Remembering Leo Breiman.....	ADELE CUTLER	1621
Leo Breiman: An important intellectual and personal force in statistics, my life and that of many others .....	PETER J. BICKEL	1634
Remembrance of Leo Breiman.....	PETER BÜHLMANN	1638
Leo Breiman .....	MICHAEL I. JORDAN	1642
Remembering Leo Breiman .....	RICHARD A. OLSHEN	1644
Remembering Leo .....	JEROME H. FRIEDMAN	1649
Selected recollections of my relationship with Leo Breiman .....	CHARLES J. STONE	1652
Leo and me .....	JACOB FELDMAN	1656
Remembering Leo .....	BIN YU	1657

# 2019年度リーディングDAT(Data Analytics Talents)

新型コロナウイルス感染症の我が国における拡大状況に鑑み、協議の結果、本講座の開催を中止させていただくことになりました。

2020.2.25

## 【リーディングDAT講座】L-S 決定木とアンサンブル学習の基礎と実践

● 申込みに関するQ&A

### 内 容

#### ① 内容

現在のデータ社会では多種多様なデータの利活用が求められています。決定木アンサンブルは高速、高精度、非線形で柔軟な機械学習法の一つとしてデータサイエンティストの道具箱に定着してきました。母集団の分布や生成過程を仮定しその未知母数を標本から推定するデータモデリング型の統計的推定と異なり、決定木アンサンブルは極めてアルゴリズム的な手法です。本講義では、決定木学習の多様な背景・歴史、その仕組みと利点・欠点の整理から始めて、決定木を基底とするアンサンブル学習の基礎と実践について、関連手法のライブラリの紹介を含め勘所を1日で講義します。

※2019年度L-B2講座内の「決定木に基づくアンサンブル学習(講師：瀧川一学)」と30%程度重複します。

※時間割は[こちら \(PDF\)](#)

#### ② キーワード

分類木・回帰木・モデル木、プロダクションルールと論理推論、CART、C4.5とM5、バギングとブースティング、ランダム部分空間法、ランダムフォレスト法、ランダム木とExtraTrees、勾配ブースティング法、確率的勾配降下、XGBoostとLightGBM

#### ③ 受講者に期待する予備知識やレベル

簡単な微積分・行列計算・確率の計算の知識は前提とします。

#### ④ 参考書

適宜講義内で紹介します。

### 講 師

瀧川一学（理化学研究所、北海道大学）

#### ⑤ 講師プロフィール

離散構造を伴う機械学習・データマイニング、生命科学・量子化学・材料科学でのデータ駆動型研究

※やむを得ない事情により、断りなく講師が変更となる可能性があります。

### 日 時

2020年3月3日(火)10時～17時 (開場9時30分)

### 申込受付期間

2020年1月6日(月)10時～1月20日(月)10時 **申込みの受付は終了しました。**

# 2021年度リーディングDAT(Data Analytics Talents)

## 【リーディングDAT講座】L-B2 機械学習とデータサイエンスの現代的手法

② 申込みに関するQ&A

日 時	<p>2021年12月16日(木)～12月17日(金) 10時～16時30分 (Zoomウェビナーには9時30分からお入りいただけます)</p>
<p>内 容</p>	<p>● 内容          「機械学習概論」赤穂昭太郎（産業技術総合研究所）          「行列データ・テンソルデータの機械学習」今泉允聰（東京大学）  <span style="border: 2px solid red; padding: 2px;">「決定木に基づくアンサンブル学習」瀧川一学（理化学研究所、北海道大学）</span>          「ガウス過程の基礎と応用」持橋大地（統計数理研究所）  <a href="#">※講師プロフィールはこちら</a></p> <p>本年度より「機械学習概論」の講義を新たに加え、各内容に共通の基礎事項を提示。既に豊富に情報のあるニューラルネットワークに関する内容はあえて外し、それ以外の主要な手法から応用上重要な3つを選んで解説します。</p>
<p>● 受講者に期待する予備知識やレベル          簡単な微積分・行列計算・確率の計算の知識は前提とします。          また、回帰分析、最尤推定などの統計の初步的な知識があると理解に役立ちます。</p> <p>● 参考書          赤穂昭太郎「カーネル主成分分析」（岩波書店）          持橋大地、大羽成征「ガウス過程と機械学習（機械学習プロフェッショナルシリーズ）」（講談社）          Carl Edward Rasmussen、Christopher K. I. Williams「Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)」（The MIT Press）          Christopher Michael Bishop、「パターン認識と機械学習 上・下」（丸善出版）</p> <p>※やむを得ない事情により、断りなく講師が変更となる可能性があります。</p>	<p>2021年10月25日(月)10時～11月8日(月)10時</p>
<p>申込受付期間</p> <p>本講座の申込受付は、株式会社国際文献社のシステム上で行います。上記ボタンをクリックすると同社のサイトに移動します。（申込受付開始後はボタンが青色になり、クリックできるようになります。）</p> <p>100名（<span style="color: red;">申込多数の場合は抽選</span>）</p>	<p style="text-align: center;"><a href="#">申込み</a></p>

# 決定木に基づくアンサンブル学習

## 1. 決定木とは？

- (1) 決定木の歴史と基本
- (2) 発展：モデル木と多変量木

## 2. 決定森(決定木アンサンブル)モデル

- (1) Random Forest, Extra Trees
- (2) 勾配ブースティング木  
(GBDT, XGBoost, LightGBM)

## 3. 決定木モデルによるデータ分析

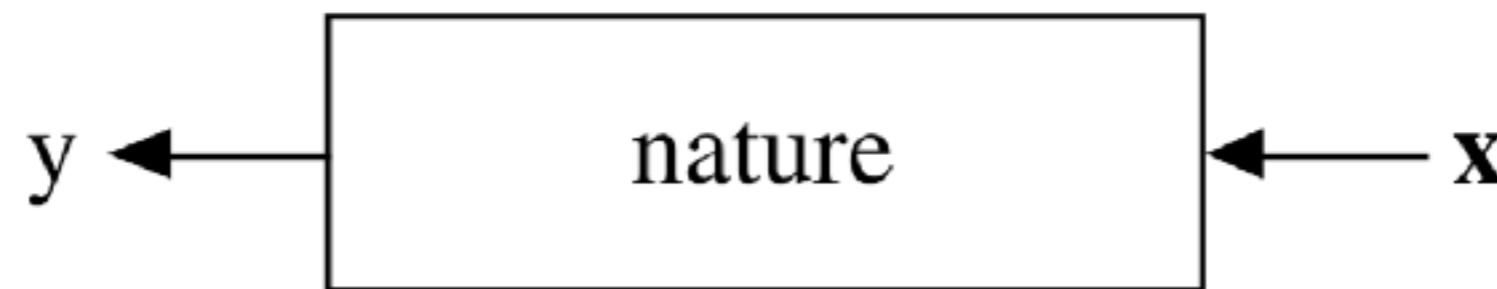
- (1) 変数重要度 (MDIとPFI), PDP, SHAP
- (2) 予測値(回帰)の信頼区間推定



瀧川一学

理化学研究所  
革新知能統合研究センター

# The Two Cultures



There are **two goals** in analyzing the data:

***Prediction.*** To be able to predict what the responses are going to be to future input variables;

***Information.*** To extract some information about how nature is associating the response variables to the input variables.

# The Two Cultures : 統計学 vs 機械學習

(伝統的な)統計学

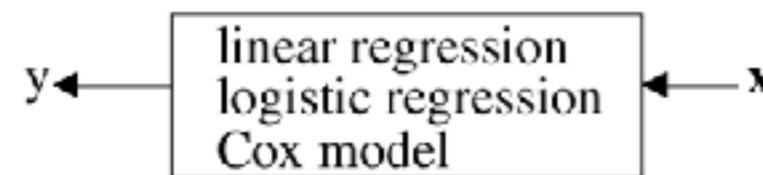


機械學習



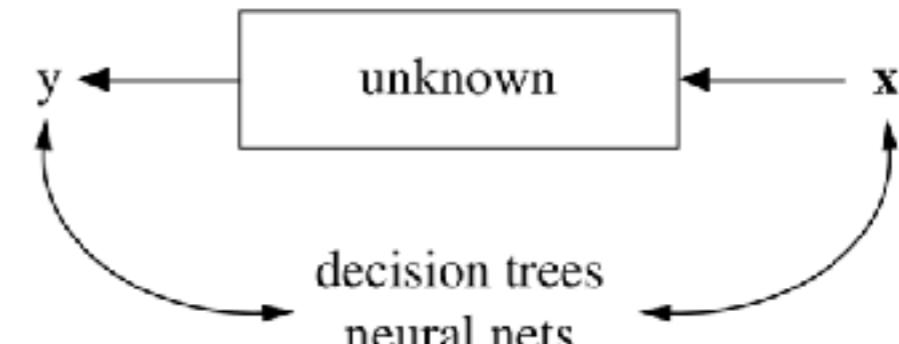
**The Data Modeling Culture**

vs **The Algorithmic Modeling Culture**



*Model validation.* Yes–no using goodness-of-fit tests and residual examination.

*Estimated culture population.* 98% of all statisticians.



*Model validation.* Measured by predictive accuracy.  
*Estimated culture population.* 2% of statisticians, many in other fields.

"Explanatory Modeling"

vs

"Predictive Modeling"

"Generative Models"

vs

"Discriminative Models"

# Leo Breimanの3レッスン

## 7.3 Recent Lessons

The advances in methodology and increases in predictive accuracy since the mid-1980s that have occurred in the research of machine learning has been phenomenal. There have been particularly exciting developments in the last five years. What has been learned? The three lessons that seem most important to one:

- Rashomon*: the multiplicity of good models;
- Occam*: the conflict between simplicity and accuracy;
- Bellman*: dimensionality—curse or blessing.

# Leo Breimanの3レッスン

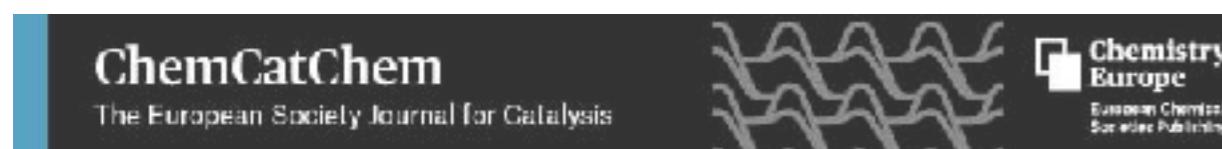
## Rashomon, Occam, Bellman

今でもこの3点はまだ色々とOpenな問題を孕んで研究されている

- "Rashomon": 良いモデルの多重性(非一意性)
  - 同程度の良い予測精度を持つ全く異なるモデルがたくさん存在する
- "Occam": モデルの解釈性と予測精度のコンフリクト
  - モデルのシンプルさ(解釈性)と予測精度の両立はとても難しい
- "Bellman": 高次元データが引き起こすメリットとデメリット
  - 高次元な表現(関係しそうなできるだけ多くの変量)を扱うべきなのか
  - 伝統的な統計学のように支配的な少数の変量を検討し分析すべきなのか
  - キッキンシンク回帰(思いつく変数全部入りモデル)・特徴量エンジニアリングと擬似相関・Rashomon効果の増大リスク

# 私たちの苦闘：不均一系触媒のデザインと探索

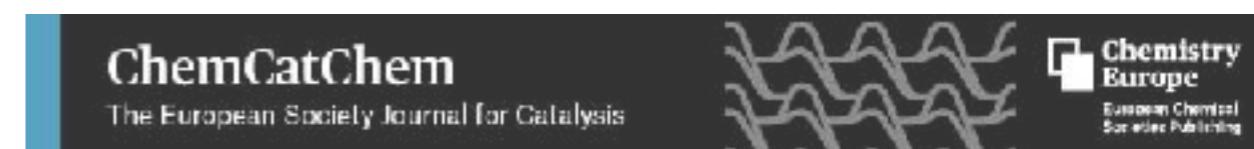
対象：工業合成・排ガス浄化・メタン転換など固体触媒表面上の気相反応



**Analysis of Updated Literature Data up to 2019 on the Oxidative Coupling of Methane Using an Extrapolative Machine-Learning Method to Identify Novel Catalysts**

Dr. Shinya Mine, Motoshi Takao, Taichi Yamaguchi, Dr. Takashi Toyao, Dr. Zen Maeno, Dr. S. M. A. Hakim Siddiki, Dr. Satoru Takakusagi, Prof. Ken-ichi Shimizu, Dr. Ichigaku Takigawa

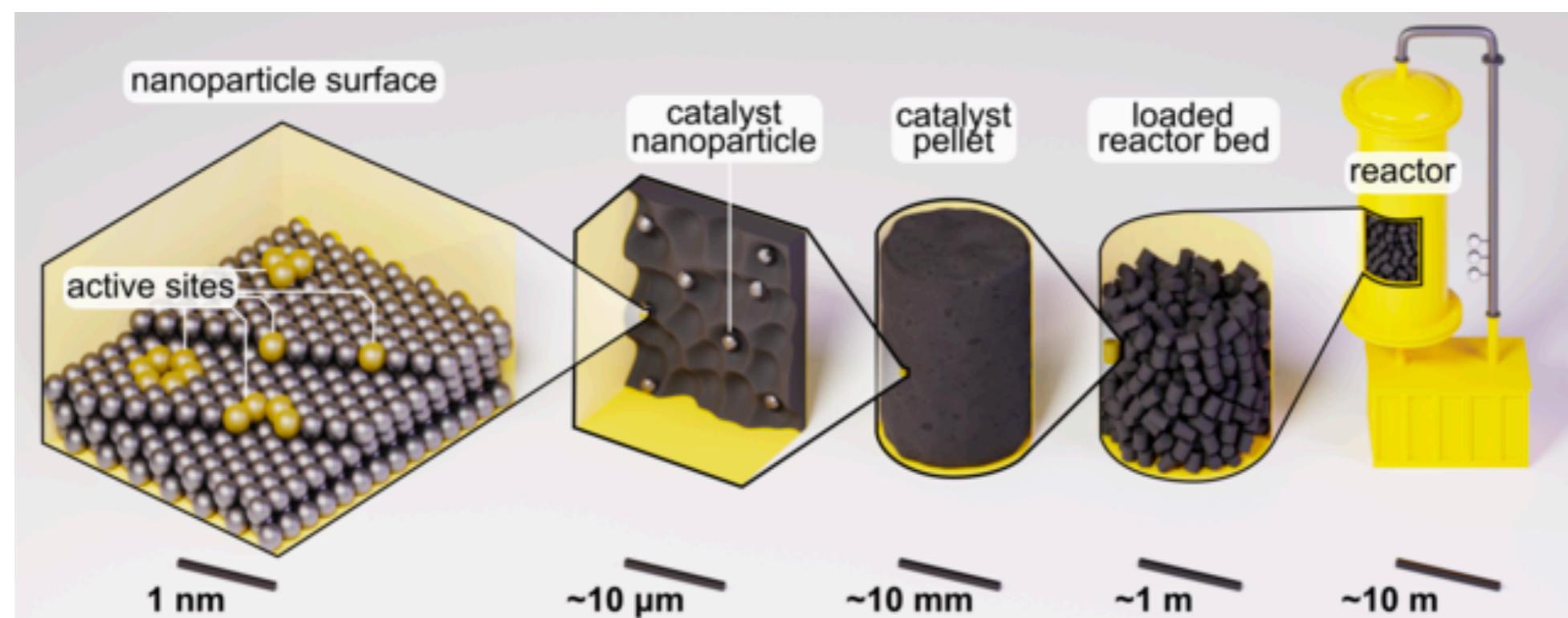
First published: 31 May 2021 | <https://doi.org/10.1002/cctc.202100495>



**Statistical Analysis and Discovery of Heterogeneous Catalysts Based on Machine Learning from Diverse Published Data**

Keisuke Suzuki, Dr. Takashi Toyao, Dr. Zen Maeno, Dr. Satoru Takakusagi, Prof. Ken-ichi Shimizu, Dr. Ichigaku Takigawa

First published: 09 July 2019 | <https://doi.org/10.1002/cctc.201900971> | Citations: 10



# 不均一系触媒研究での機械学習の活用

JST CREST 革新材料開発 (細野領域)

触媒インフォマティクスの創成のための実験・理論・データ科学的研究



北海道大学 触媒科学研究所  
Hokkaido University, Institute for Catalysis



清水研一



高草木 達



鳥屋尾 隆



前野 禅

高尾基史, 峯 真也, 鈴木慶介

- Mine+ *ChemCatChem.* 2021.
- Toyao+, *ACS Catalysis.* 2020. (Review)
- Liu+, *The Journal of Physical Chemistry C.* 2020.
- Suzuki+ *ChemCatChem.* 2019. (Front Cover)
- Kamachi+ *The Journal of Physical Chemistry C.* 2019.
- Hinuma+ *The Journal of Physical Chemistry C.* 2018.
- Toyao+, *The Journal of Physical Chemistry C.* 2018
- Takigawa+ *RSC Advances.* 2016.

この続き物の  
研究を紹介

# 参考 : Toyao+, ACS Catalysis. 2020. (Review)



Review

Cite This: ACS Catal. 2020, 10, 2260–2297

pubs.acs.org/acscatalysis

## Machine Learning for Catalysis Informatics: Recent Applications and Prospects

Takashi Toyao,<sup>†,‡,§,||,¶,ID</sup> Zen Maeno,<sup>†</sup> Satoru Takakusagi,<sup>†</sup> Takashi Kamachi,<sup>‡,§,ID</sup> Ichigaku Takigawa,<sup>\*,||,¶,ID</sup> and Ken-ichi Shimizu<sup>\*,†,‡,§,ID</sup>

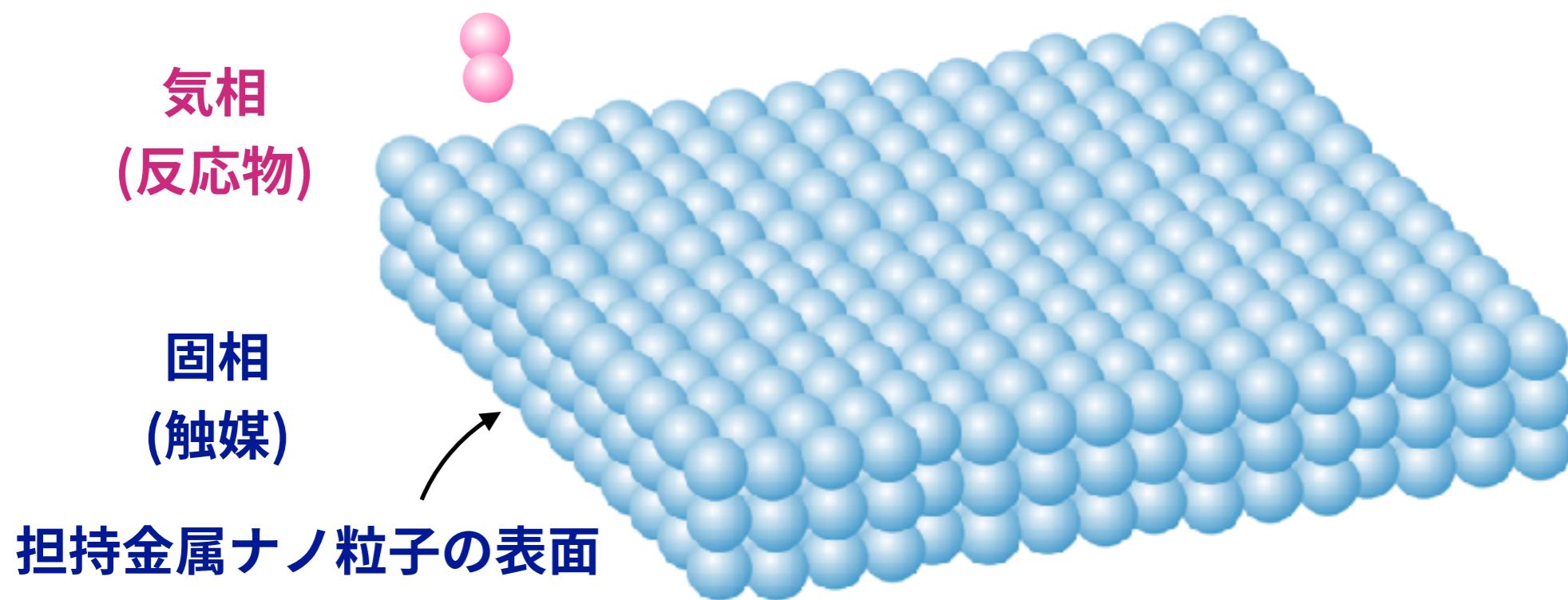
### Review Comments

- *This is an excellent review on a very timely subject, which is highly suitable for ACS Catalysis. ... I don't usually recommend that papers should be accepted "as is", but in this case I don't see the need for changes.*
- *I will certainly recommend it to my group and my students when it is published.*
- *The manuscript gives an excellent overview in the field of machine learning especially with regard to heterogeneous catalysis and I would highly recommend the article for the publication in ACS Catalysis.*
- *This is one of the best reviews for catalyst informatics that the reviewer has read. In particular, the chapter 2 delivers a very good tutorial, which is concisely and professionally written.*

2章が機械学習のユーザガイド(数式なし)になっています！

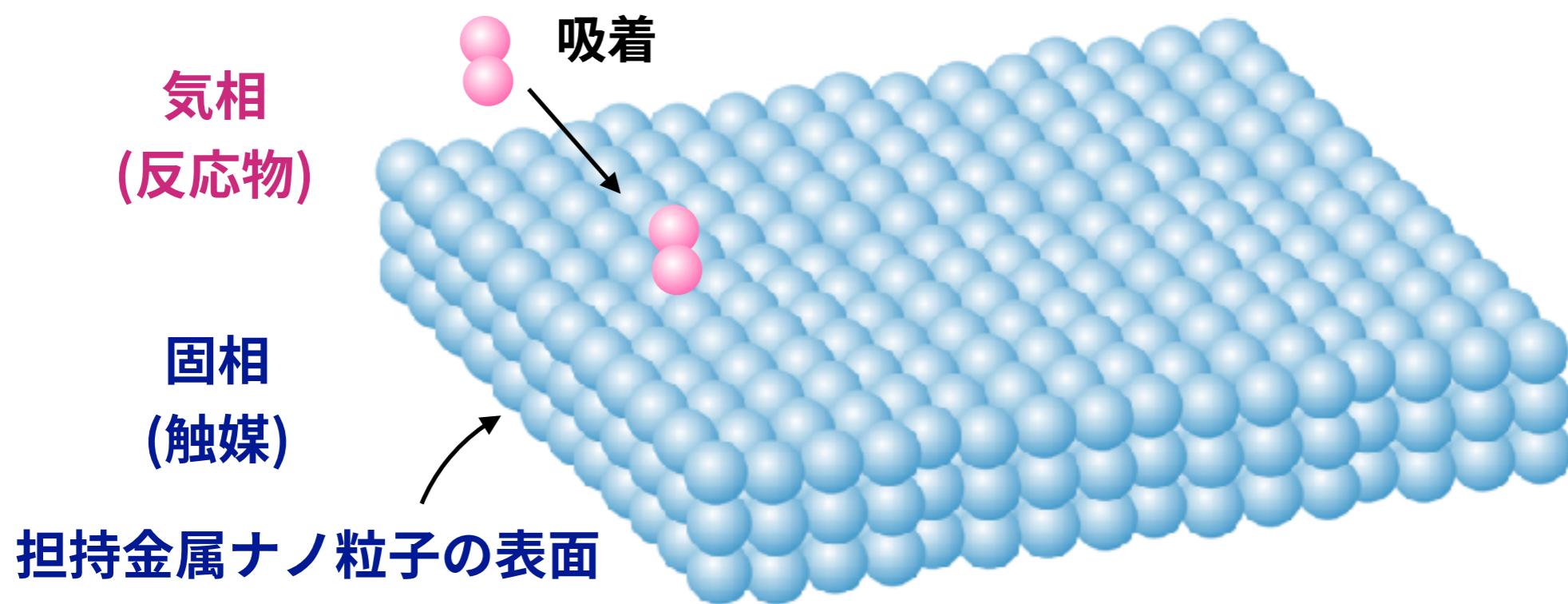
# 問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



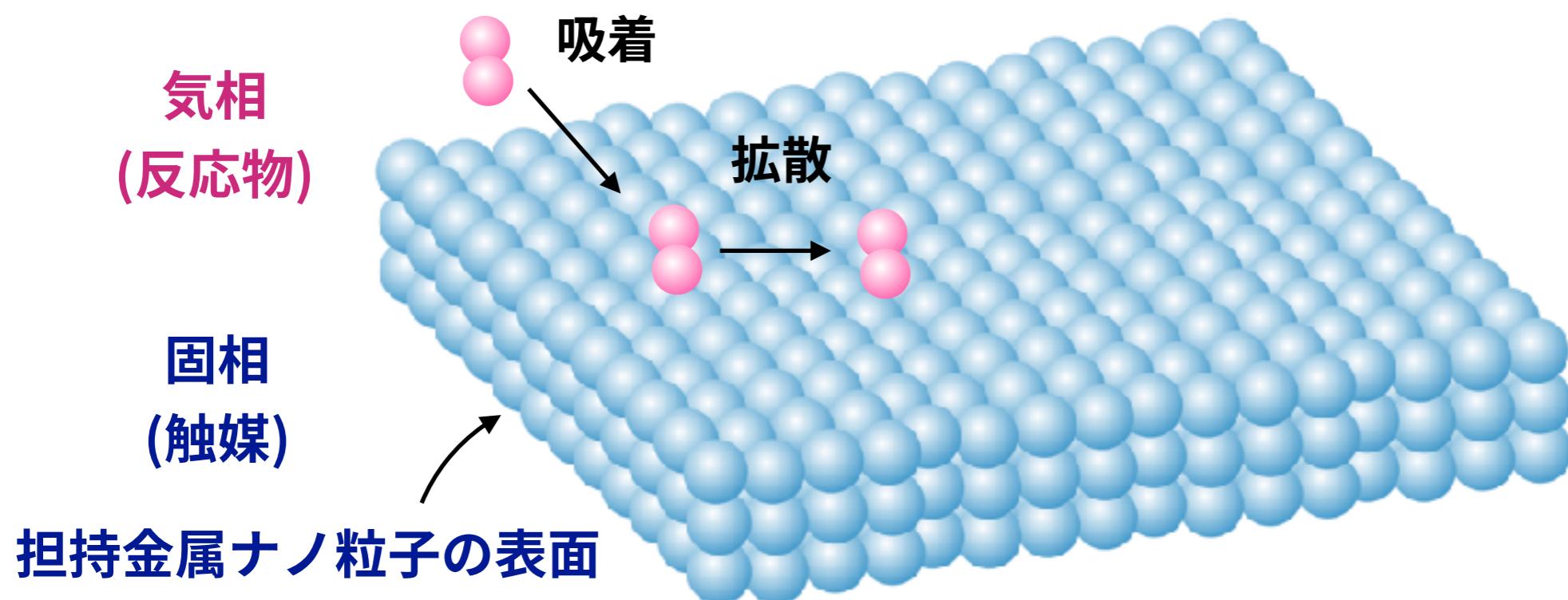
# 問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



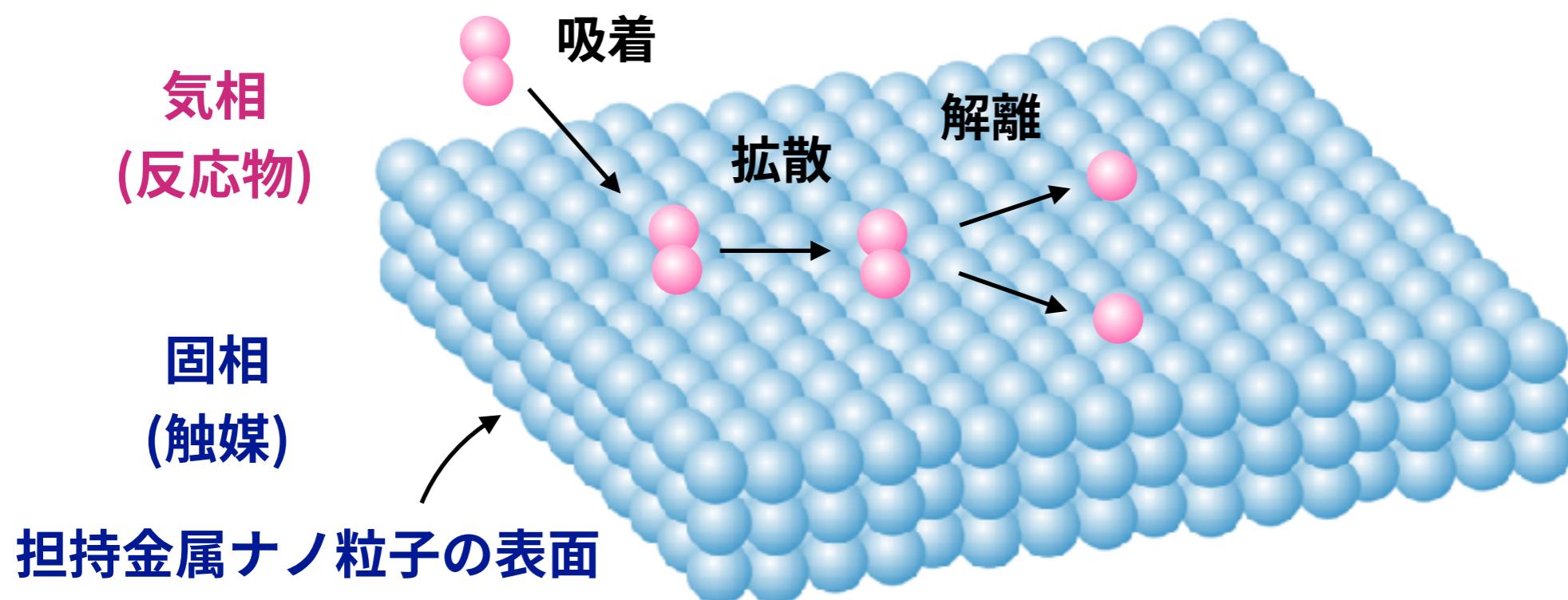
# 問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



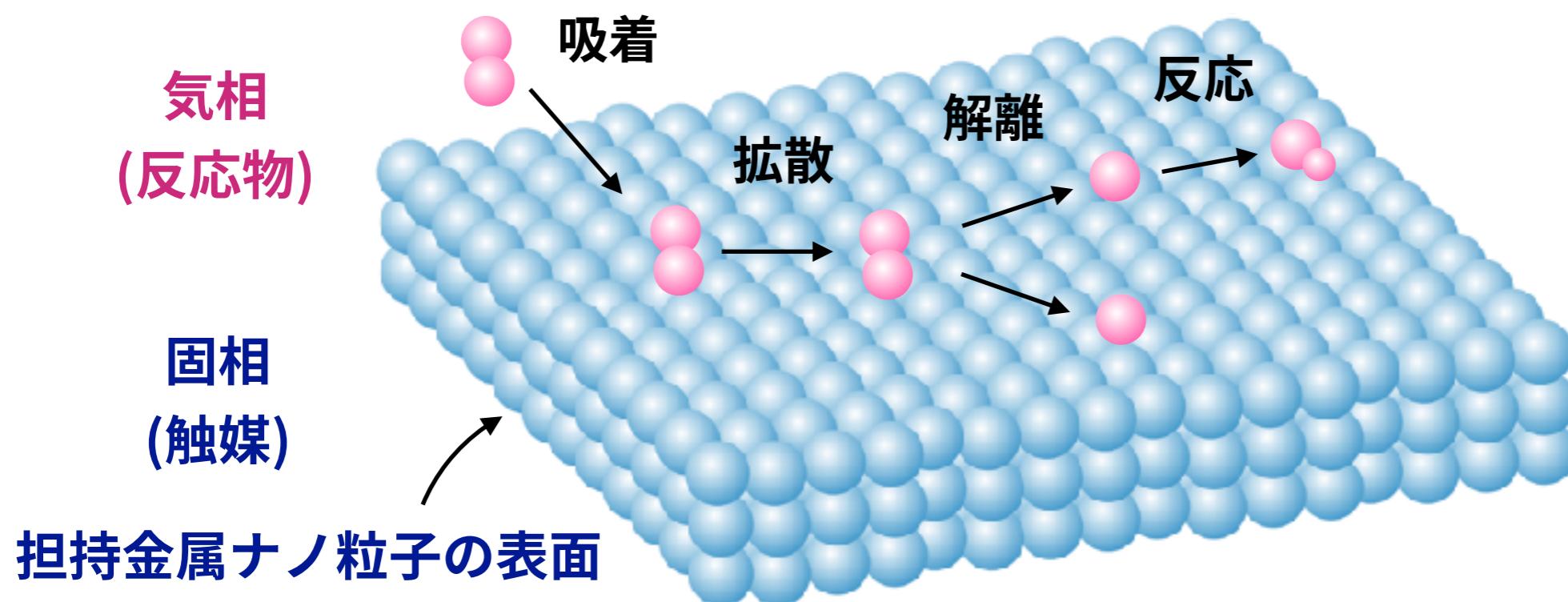
# 問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



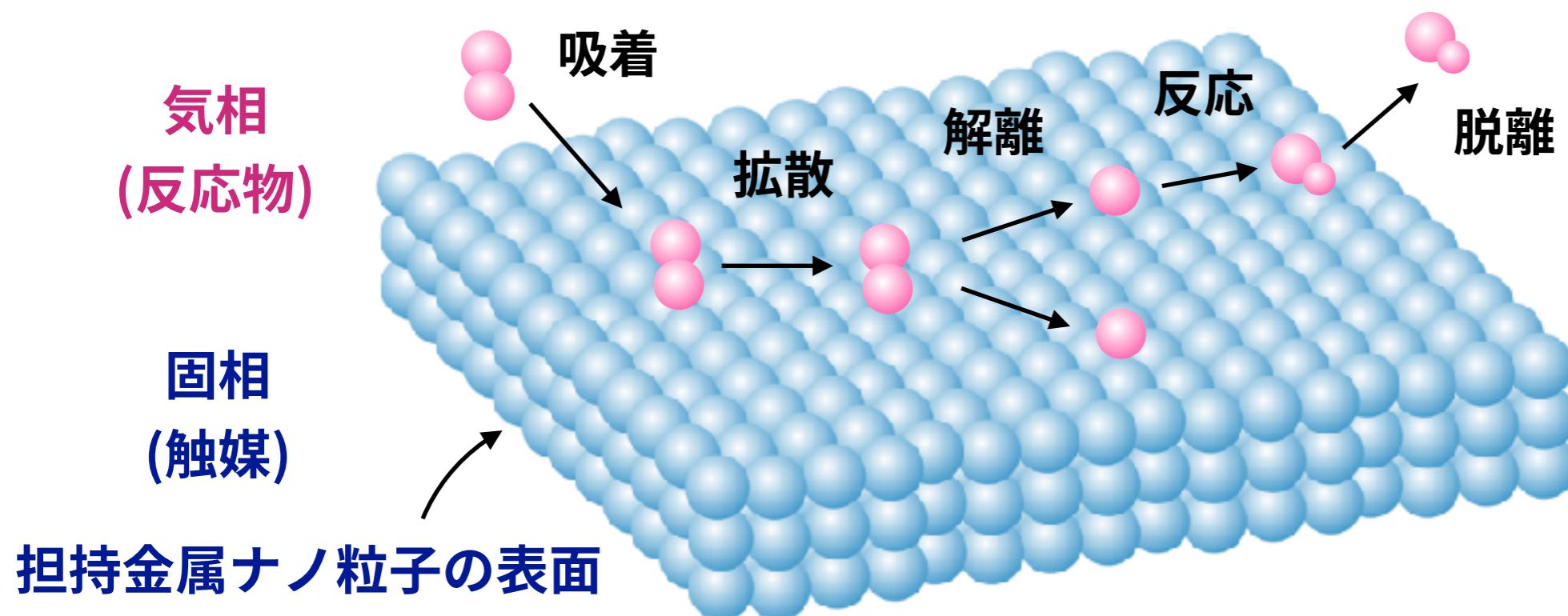
# 問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



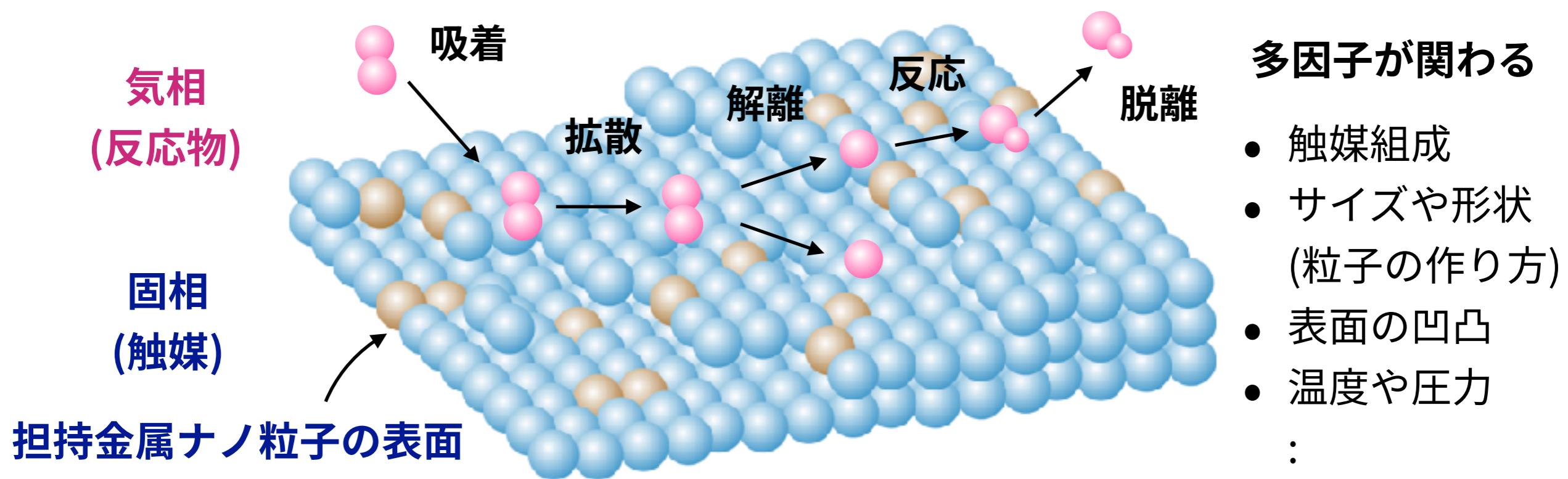
# 問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



# 問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



# 問題の難しさの確認

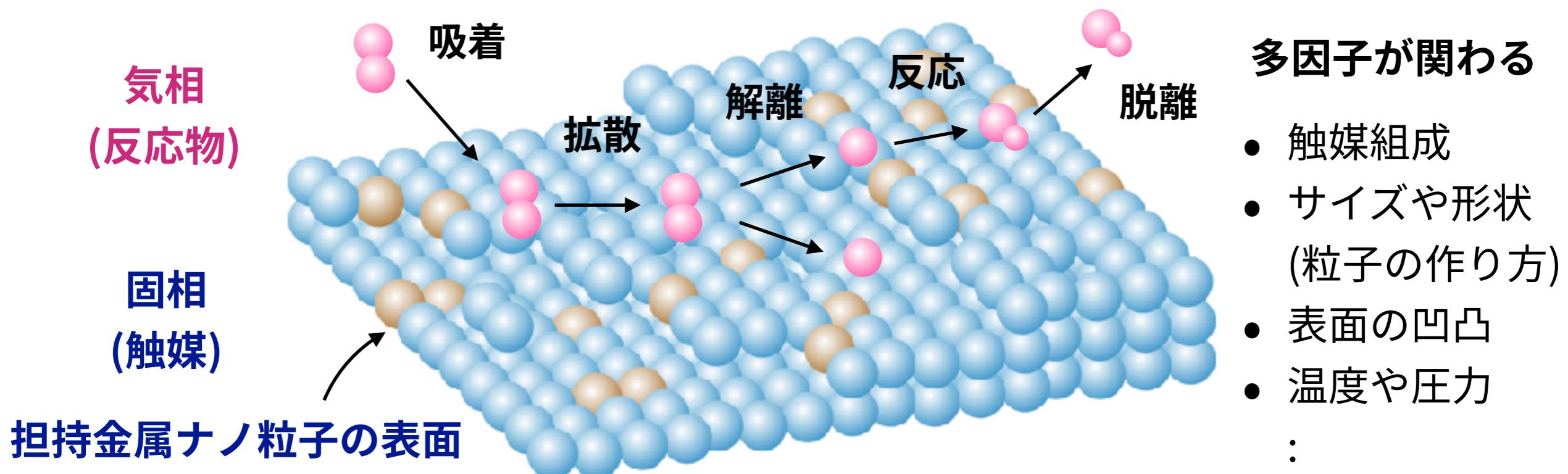
「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ

→ そもそも「表面」が悪魔的な難しさ (“表面科学”)



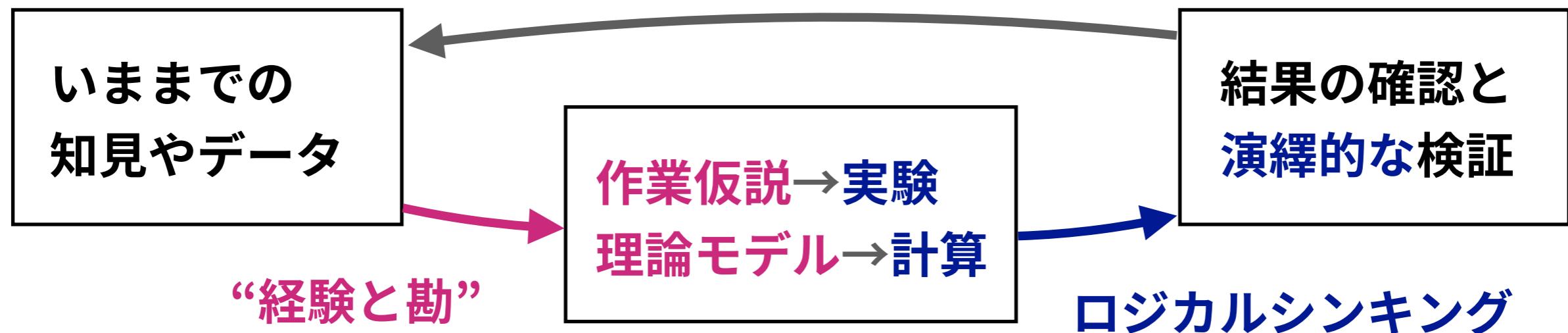
God made the bulk;  
the **surface** was invented by the devil

パウリ大先生



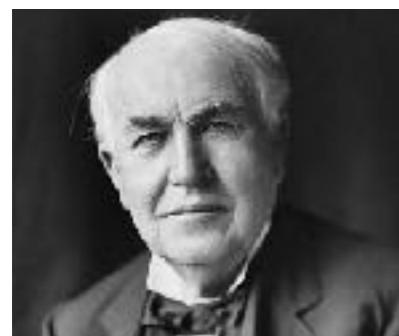
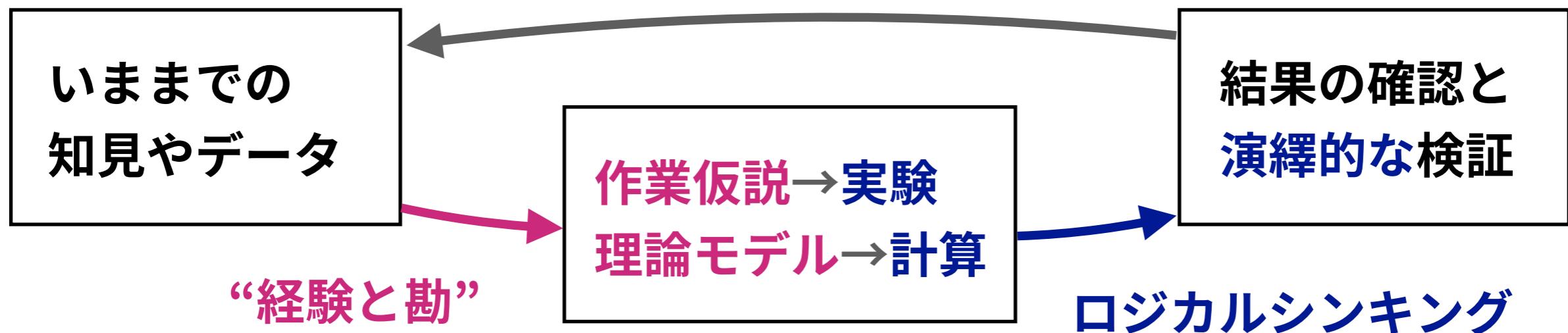
# 実験が主導して良い触媒が見つかってきたことは驚異的

## 仮説演繹法



# 実験が主導して良い触媒が見つかってきたことは驚異的

## 仮説演繹法 or “エジソン的な”経験論

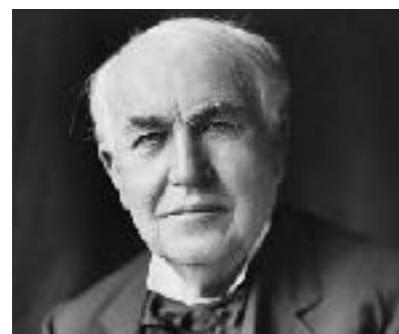
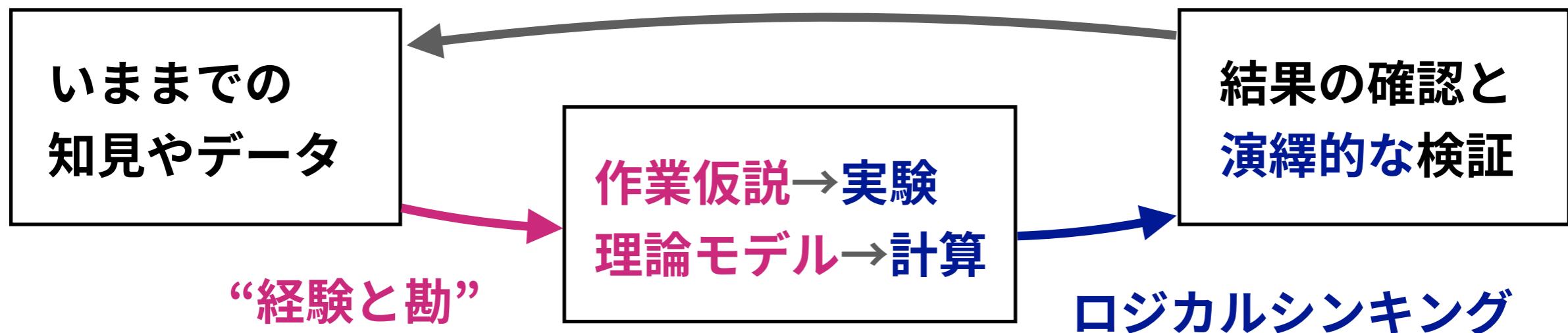


エジソン大先生

- Genius is 1% inspiration and 99% perspiration.
- There is no substitute for hard work.
- I have not failed. I've just found 10,000 ways that won't work.

# 実験が主導して良い触媒が見つかってきたことは驚異的

## 仮説演繹法 or “エジソン的な”経験論



エジソン大先生

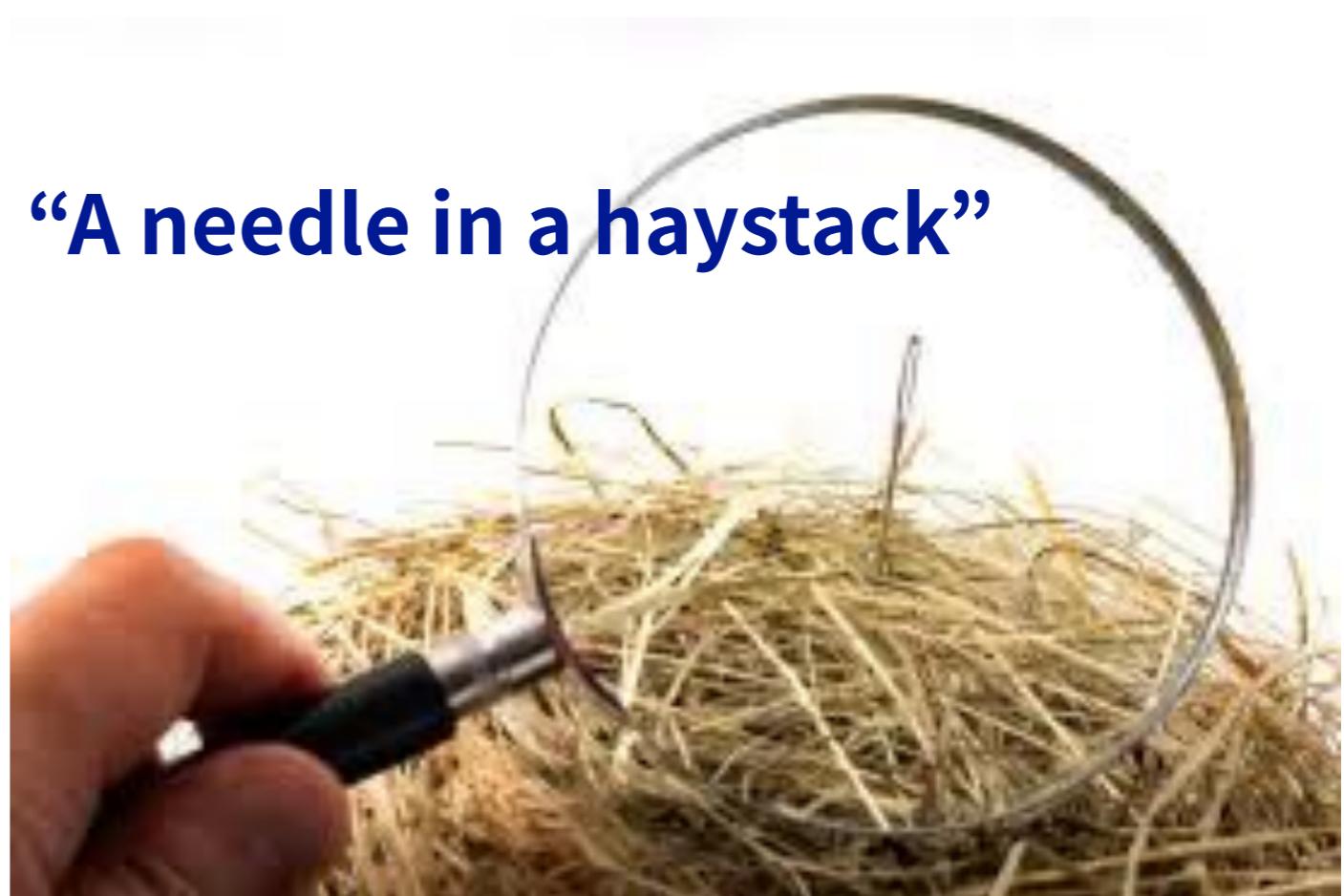
- Genius is 1% inspiration and 99% perspiration.
- There is no substitute for hard work.
- I have not failed. I've just found 10,000 ways that won't work.

要約すると「努力あるのみ！とにかくたくさんがんばれ！」と言っている。  
→ お金の投入 + 人海戦術(=ポスドクや学生の過酷な労働?)でGO

# それでも！「発見」はものすごくイベントである

想定できる「触媒+実験条件+方法」の数は**天文学的に巨大**

- 😭 有限の時間・コストを生きる私たちが試せるのはほんの一部
- 😭 複雑化するニーズを反映した素晴らしい画期的な触媒が  
見つかる確率は理屈上は絶望的に低い…はず



**“A needle in a haystack”**

# つまり「セレンディピティ ≠ 偶然の幸運」？

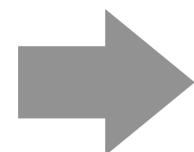
幸運は準備されたものにだけ降りる！！

- 何を実験するか(仮説形成)はふつう完全にランダム(いきあたりばったり)ではない。
- 経験と勘：「研究者のセンス」や「腕の見せ所」
- 優れた実験科学者の「勘ピュータ(経験と勘)」はランダムではなく何らかの指向性を持つ

# つまり「セレンディピティ ≠ 偶然の幸運」？

幸運は準備されたものにだけ降りる！！

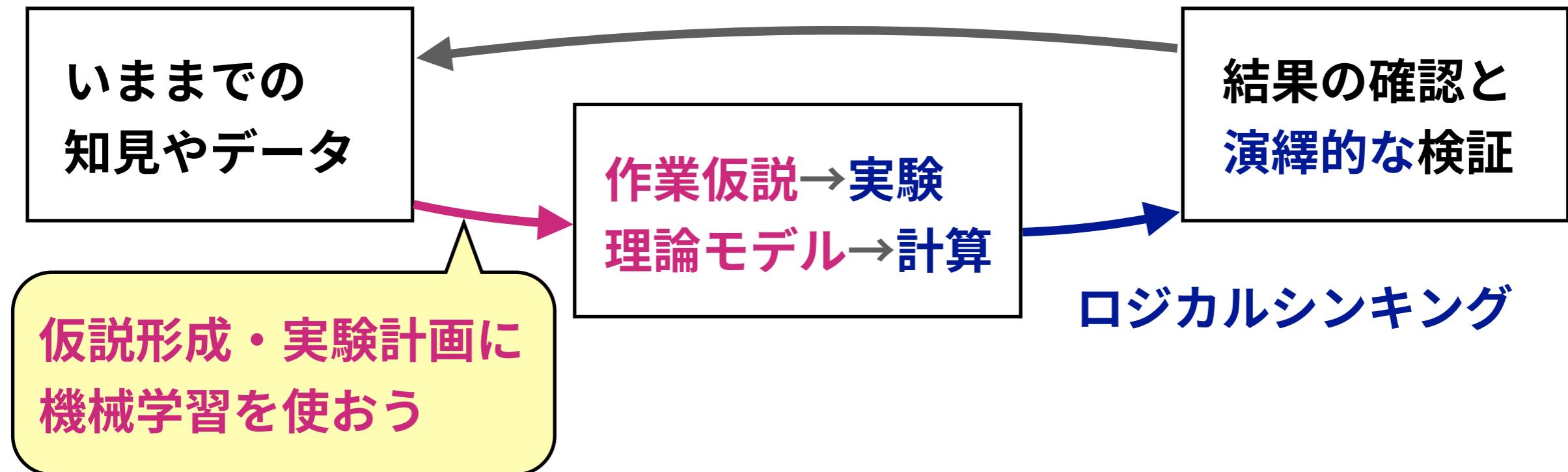
- 何を実験するか(仮説形成)はふつう完全にランダム(いきあたりばったり)ではない。
- 経験と勘：「研究者のセンス」や「腕の見せ所」
- 優れた実験科学者の「勘ピュータ(経験と勘)」はランダムではなく何らかの指向性を持つ



このあたりに“**data-driven**”が貢献できる余地がある

問題：実際にはデータ化できない情報がほとんどなので  
データ化できる情報の中の「どういうデータでdriveするのか」

# どういうデータで仮説形成・実験計画をdriveできるか?



- 優れた実験科学者に今までの全人生で入力された情報は膨大 (データ化されない情報がほとんど)
- データ駆動：手に入るデータから近似的に迫るしかないが 人間の認知限界の制約や思い込みによる束縛から **自由**になる  
⇄ 人間は多数の因子の複雑な多次元相関を把握できない

# 機械学習をどう活用するか

機械学習を活用するためには

**機械学習モデルを訓練するための「データ」をどうするかが鍵**

- 文献から集めた実際の実験データ報告を使う
- ラボで実験して蓄積したデータを使う
- シミュレーション(計算化学)で蓄積したデータを使う
- 上記3つ全部使う
- 実験計測機器にセンサーをつけまくり、実験しているところもビデオ録画し、実験者の頭にもカメラつけて実験者視野もビデオ録画し、実験者の体に動作センサつけて記録し、実験ノートもスキャンし、あらゆる関連論文や教科書も全部電子化し、…

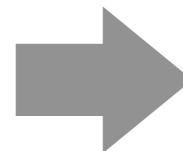
# 文献から集めた実際の実験データ報告を使う

計算化学は飛躍的発展を遂げているが理論と現実のギャップはいつまでも存在する。

現実をすべてモデル化することは不可能

→ その定義からして“モデル”とは何らかの捨象や近似を含む。

- 実験条件やプロセス条件などの因子
- 理論の際に諦めた(or ざっくり近似した)細かすぎる無数の要因
- 1モルにはアボガドロ定数( $6 \times 10^{23}$ )個の要素が本当はある



**文献から集めた実際の実験データ報告を使う**

「過去に報告された現実を見てみよう」

# 文献から集めた実際の実験データ報告を使う

対象はメタンの酸化カップリング反応、目的変数はC<sub>2</sub>収率

従来研究(Zavyalova et al, 2011)による2010年以前の **1868例** に  
2010~2020年の新たな例を加え **4759例** にまで拡充！



CCM

	A	B	C	N	O	R	S	T	Y	Z	RA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM		
	Cation 1	Cation 1	Anion 1	Anion 1	Promotor	Support	Support	Preparati	Temperat	p(CH <sub>4</sub> )	p(O <sub>2</sub> )	p(CH <sub>4</sub> )/p(O <sub>2</sub> )	p total,	Contact time, s	X(O <sub>2</sub> ) %	X(CH <sub>4</sub> ) %	S(CO) %	S(C <sub>2</sub> ) %	S(C <sub>2</sub> ) %	S(C <sub>2</sub> ) %	Y(C <sub>2</sub> ) %				
	public	mol%	mol%	mol%		1	1 mol%	on	ure, °C	bar	bar	bar	bar	bar											
1	1 Mn	9.2				All	90.8	Impregnat	1073	0.40	0.08	4.8	1.0		0.04	11.0					45.5	5.0			
20	3 Li	30.3				n.e.		n.e.	993	0.08	0.04	2.0	1.0		5.20	85.0	38.0					50.0	19.0		
21	4 Mg	66.7	S	35.3				Impregnat	1019	0.85	0.08	8.1	1.0		1.40	39.0	4.0					23.0	41.0	64.0	2.8
22	4 Mg	66.0	S	46.0				Impregnat	1017	0.86	0.08	8.3	1.0		3.00	65.0	10.0					27.0	40.0	67.0	6.7
23	4 Na	7.0	S	60.0				Impregnat	1017	0.84	0.08	8.0	1.0		0.19	35.0	3.0					23.0	19.0	42.0	1.3
75	6 Po	20.0				All	80.0	n.a.	1030	0.96	0.05	19.2	1.0		0.40	100.0	8.8					17.6	32.8	50.4	3.4
76	6 Po	20.0				Si	80.0	n.a.	1103	0.96	0.05	19.2	1.0		0.56	44.1	18.7					18.7	20.6	39.2	7.3
483	116 K	3.0	G	3.0	Cl	All	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		30.8					33.6	10.3		
487	116 Li	3.0	Cl	3.0	Cl	All	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		2.2					76.9	1.7		
488	116 Ba	3.0	G	6.0	Cl	All	82.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		32.1					32.1	10.3		
489	116 Na	3.0	G	3.0	Cl	All	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		35.0					33.5	11.7		
490	116 Cs	3.0	Cl	3.0	Cl	All	86.6	Impregnat	973	0.10	0.05	2.0	1.0		1.50		30.2					24.3	7.3		
491	116 Ag	16.0			Cl	All	82.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		20.4					0.0	0.0		
492	116 Ag	16.0	C	41.0	Cl			Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.8					0.5	0.1		
493	116 Pr	6.0			Cl	All	86.6	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.8					0.2	0.1		
494	116 Pr	1.0			Cl	All	90.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.1					0.0	0.0		
495	116 Si	1.0			Cl	All	81.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		23.4					1.3	0.3		
496	116 Ba	1.0			Cl	All	81.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		27.8					0.7	0.2		
497	116 Ba	5.0			Cl	All	77.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.0					1.7	0.4		
498	116 K	3.0			Cl	All	79.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		17.2					23.3	4.0		
499	116 Ba	3.0	G	6.0	Cl	All	82.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		15.8					28.0	4.4		
500	116 Ba	3.0	C	6.0	Cl	All	73.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		27.1					30.4	8.2		
501	116 Cs	1.0	G	2.0	Cl	All	79.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		16.8					26.4	4.0		
502	116 Ag	16.0			Cl	All	82.0	Thermal	973	0.10	0.05	2.0	1.0		1.50		5.0					0.0	0.0		
503	116 Ba	3.0	Cl	6.0	Cl	All	73.0	Thermal	973	0.10	0.05	2.0	1.0		1.50		17.2					25.4	4.4		
504	116 Ba	3.0	G	6.0	Cl			Thermal	973	0.10	0.05	2.0	1.0		1.50		26.7					16.3	4.1		
505	116 Ba	3.0	G	6.0	Cl	All	73.0	Thermal	973	0.10	0.05	2.0	1.0		1.50		21.3					30.4	8.5		
506	117 Sr	3.0	Cl	6.0	Cl	All	91.0	Impregnat	1023	0.10	0.05	2.0	1.0		1.50		30.3					56.0	17.0		
507	117 Ba	26.0	C	26.0	Cl	All	44.0	Impregnat	1023	0.10	0.05	2.0	1.0		1.50		43.2					41.8	18.1		
508	117 Sr	3.0	G	6.0	Cl	All	71.0	Thermal	1013	0.05	0.05	2.0	1.0		1.50		17.0					11.0	12.5		

# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力  $x$

- 元素組成比
- 実験条件

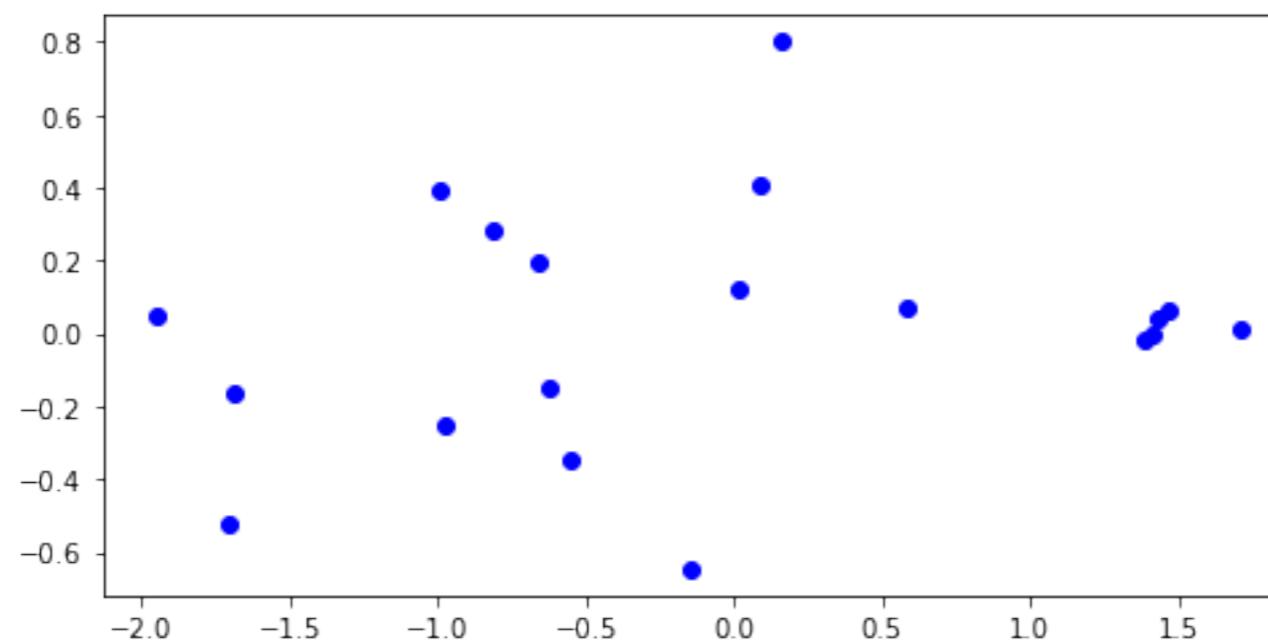
機械学習モデル

出力  $y$

- 収率

訓練データ

出力  $y$



入力  $x$

# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力  $x$

- 元素組成比
- 実験条件

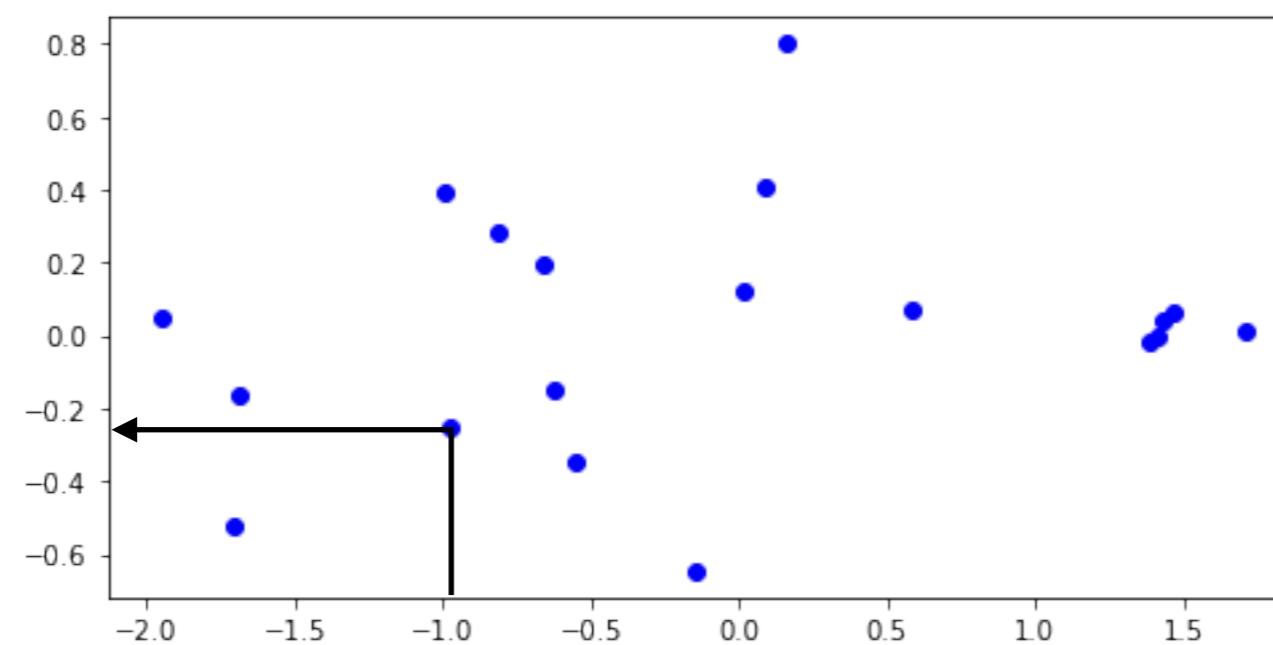
機械学習モデル

出力  $y$

- 収率

訓練データ

出力  $y$



入力  $x$

# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力  $x$

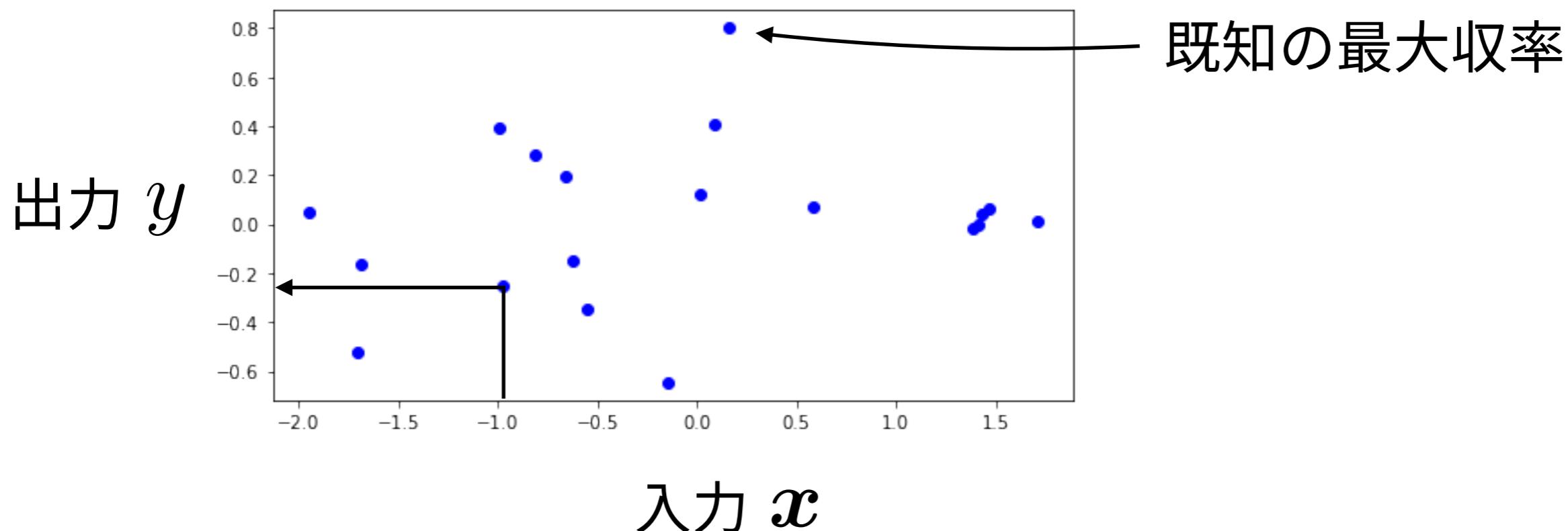
- 元素組成比
- 実験条件

機械学習モデル

出力  $y$

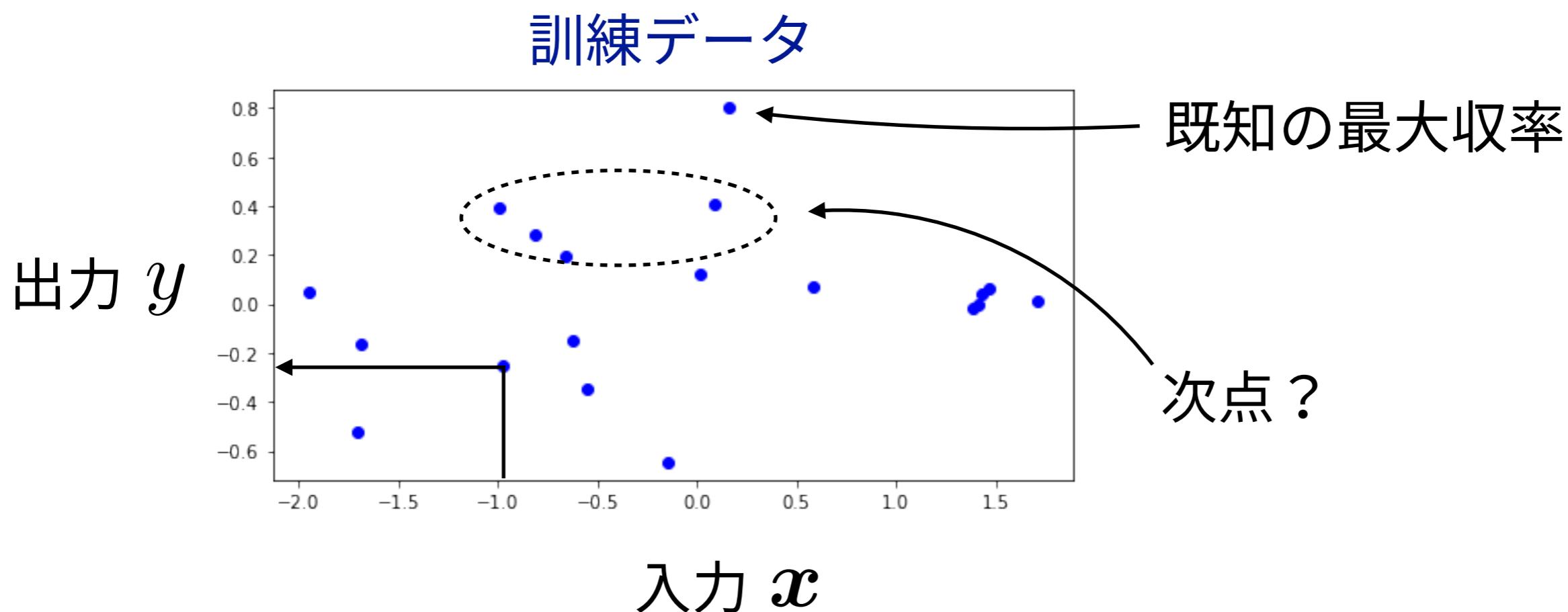
- 収率

訓練データ



# このとき機械学習に何が必要か

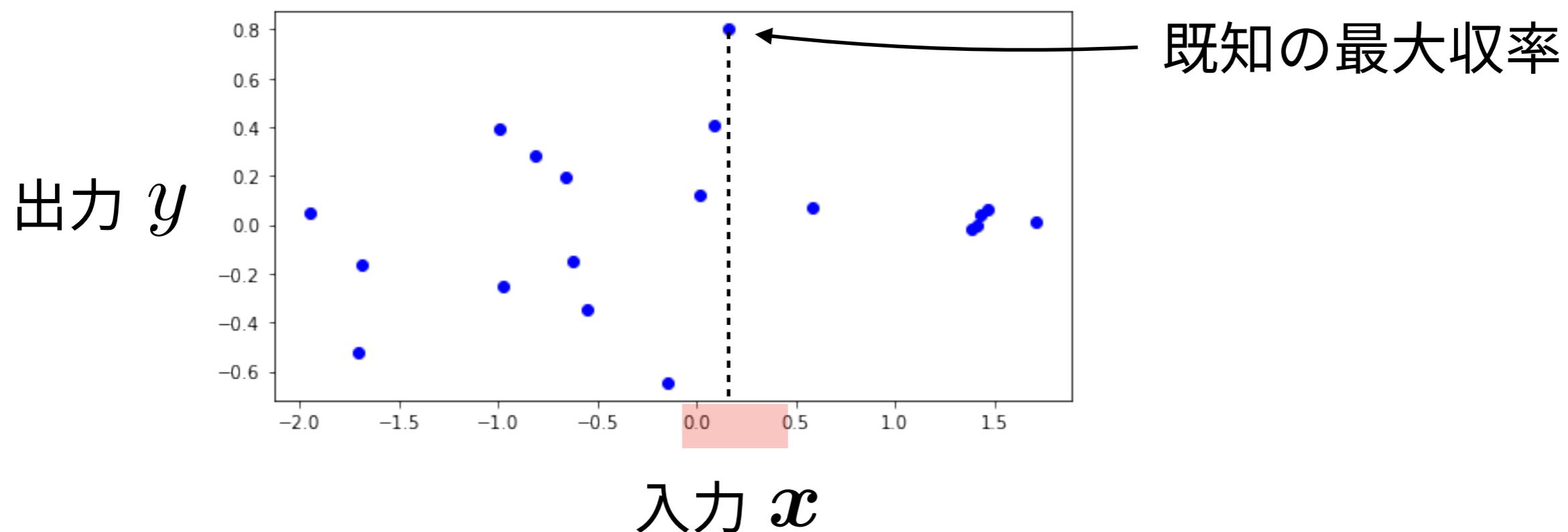
この表データで機械学習モデルを訓練し予測すれば良い？



# 知りたいことに応じて機械学習の方法や注意点が変わる

- 収率  $y$  が高い触媒と低い触媒の違いを規定する因子は何？
- 良い収率が得られる未発見の触媒  $x$  はどのあたりにある？  
(最大収率が得られた  $x$  の周辺なのだろうか？)

目的=探索 (既知の触媒より良い触媒を見つけたい！)



# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力  $x$

- 元素組成比
- 実験条件

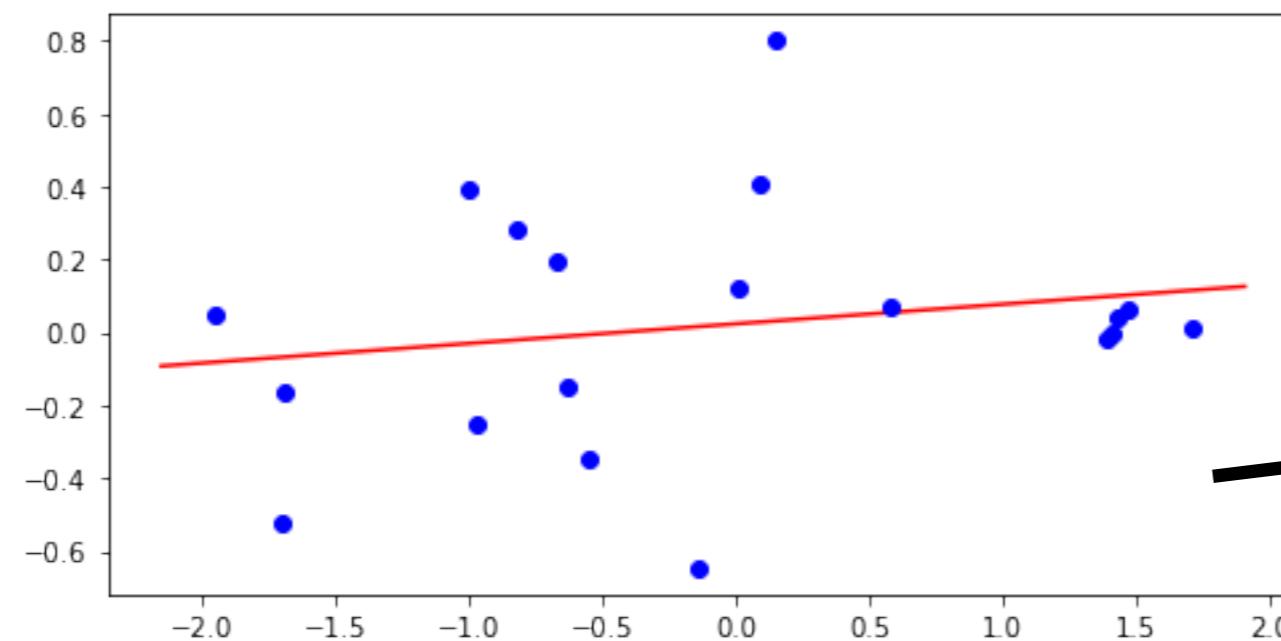
機械学習モデル

出力  $y$

- 収率

LinearRegression()

出力  $y$



入力  $x$

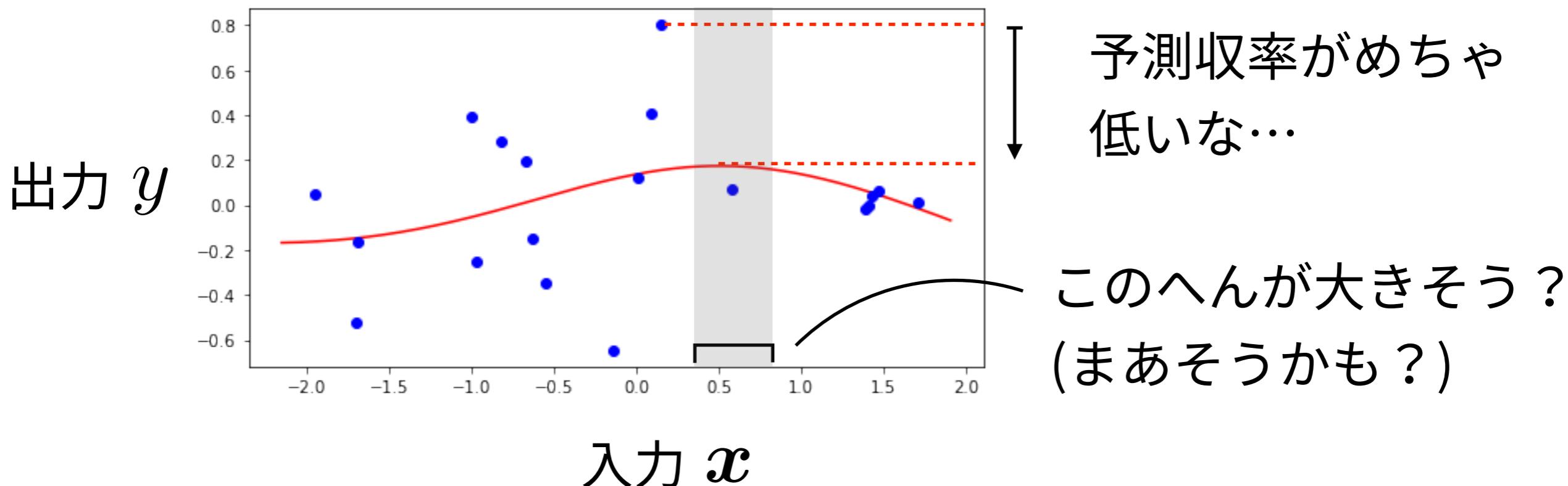
$x$  を大きくすれば  
良いという謎の  
示唆しかくれない  
(Underfit)

# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



`MLPRegressor(hidden_layer_sizes=(300, 300, 50), activation='tanh')`

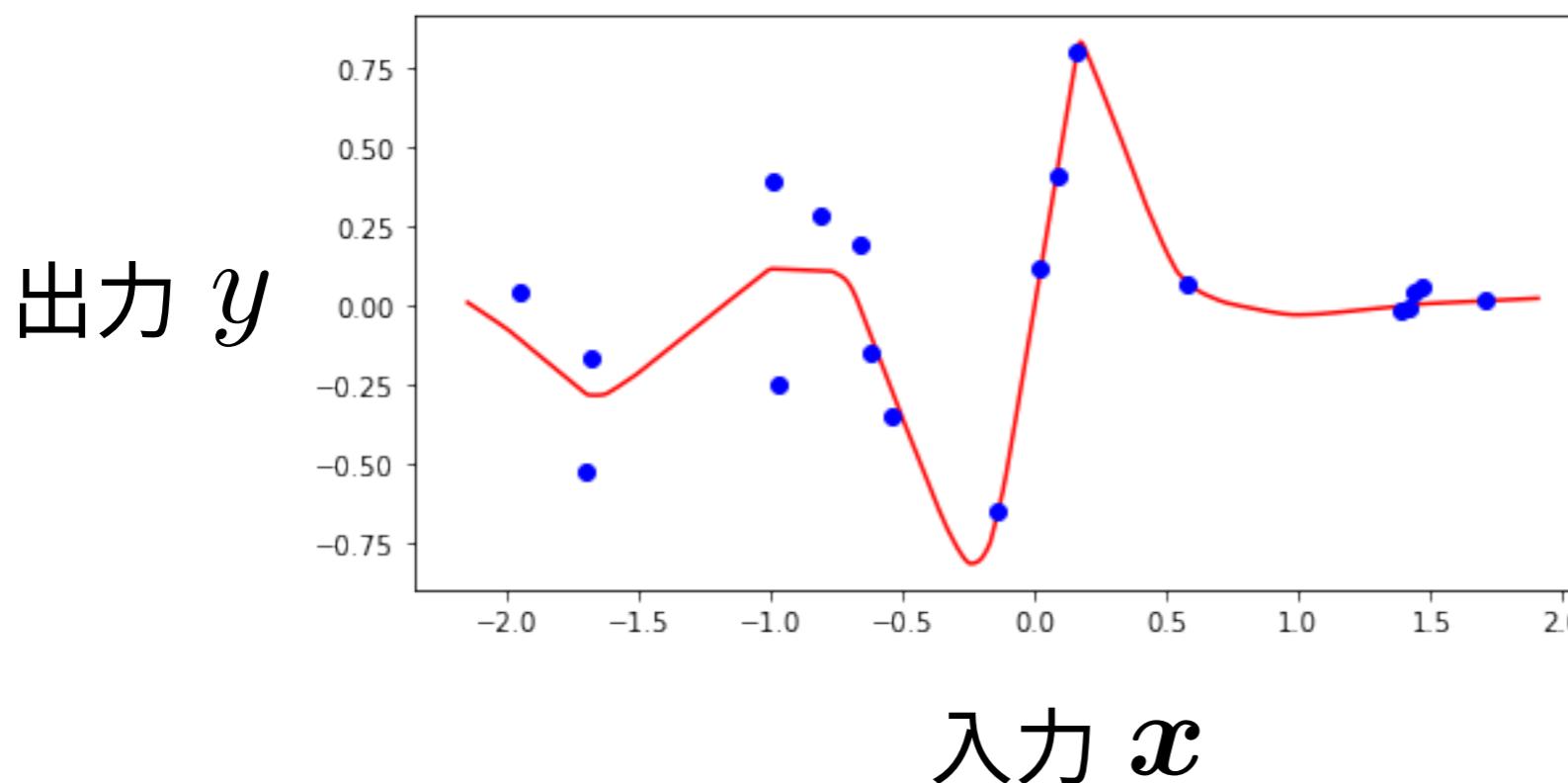


# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



`MLPRegressor(hidden_layer_sizes=(300, 300, 50), activation='relu')`



activationを「ReLU」に  
(負値をゼロ置換)

うーん？

わりとよさそうかも？

# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力  $x$

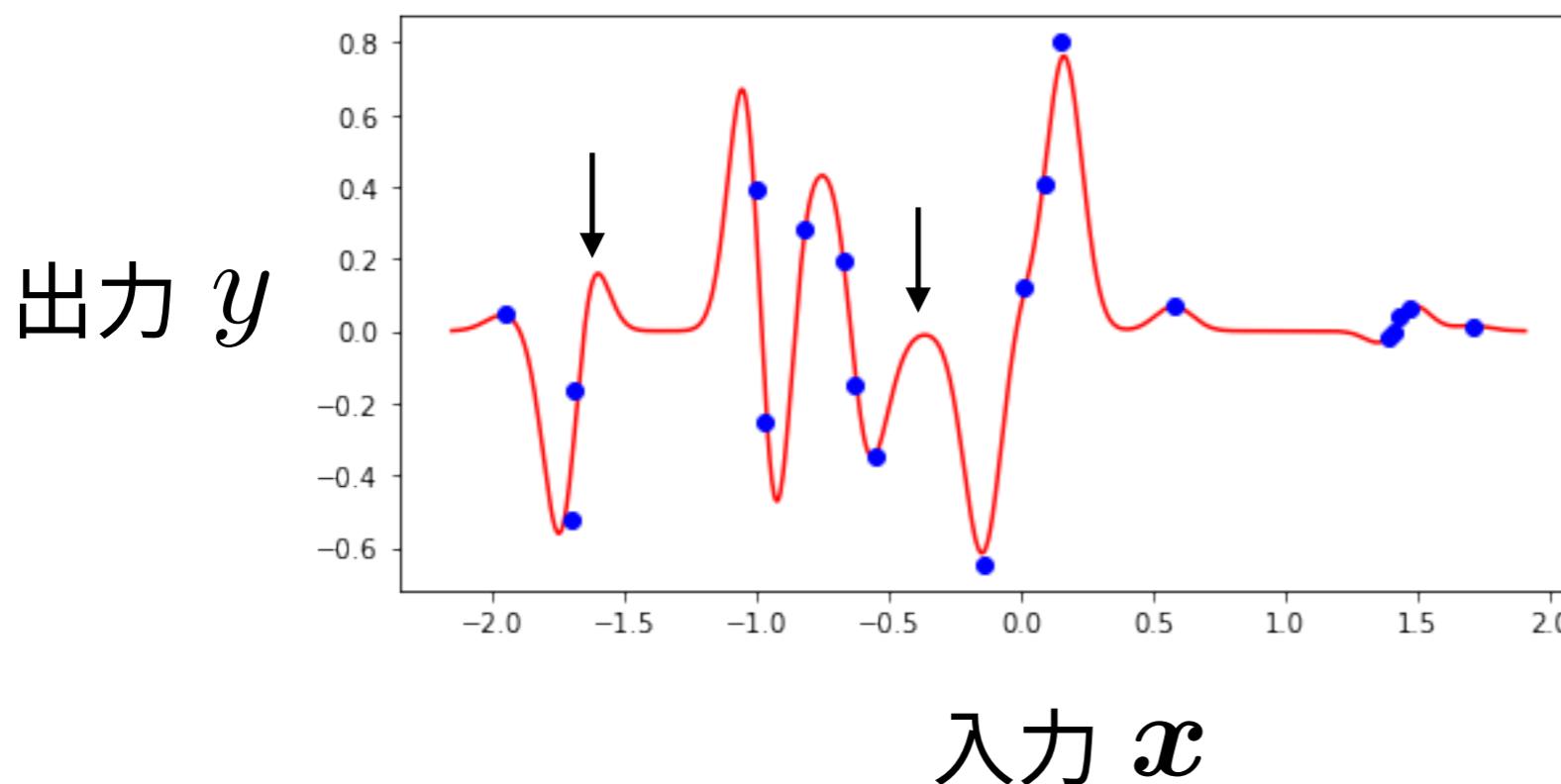
- 元素組成比
- 実験条件

機械学習モデル

出力  $y$

- 収率

`KernelRidge(kernel='rbf', gamma=100.0, alpha=0.05)`



うーん？？

特に ↓ の箇所は  
これで良いのか..！？

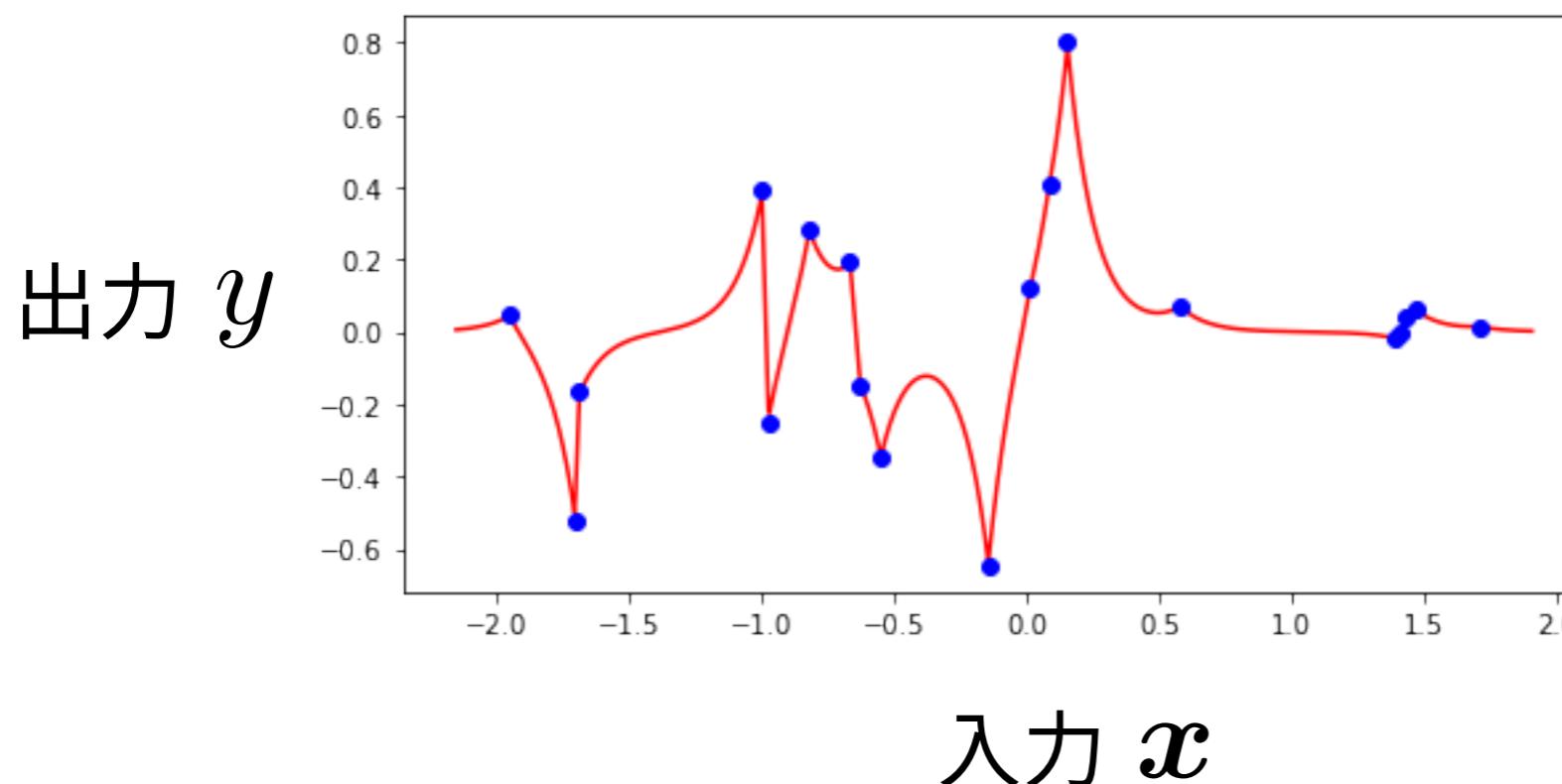
アーティファクト？

# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



`KernelRidge(kernel='laplacian', gamma=10.0, alpha=0.01)`



うーん？？

# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

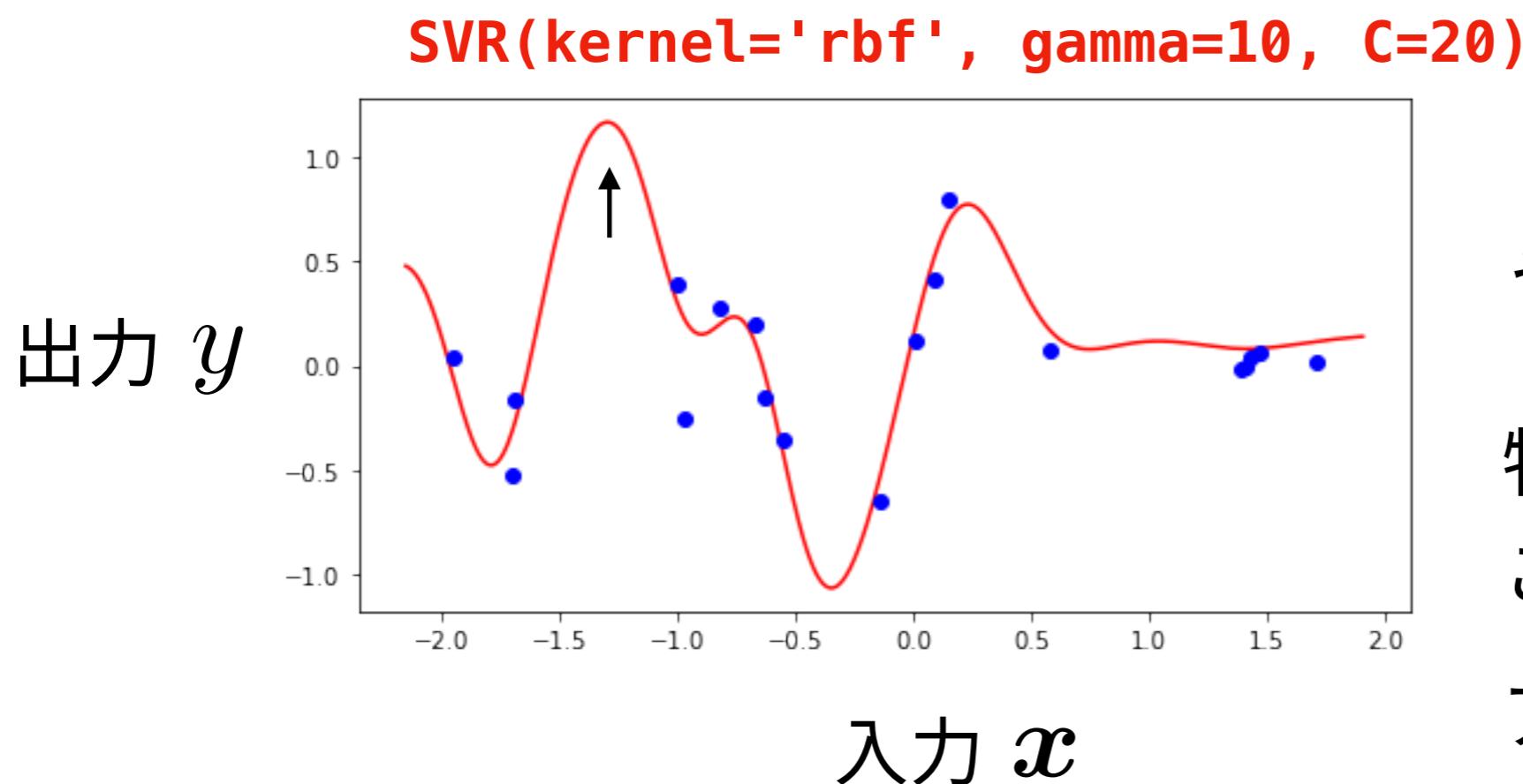
入力  $x$

- 元素組成比
- 実験条件

機械学習モデル

出力  $y$

- 収率



うーん？？？

特に ↑ の箇所は  
これで良いのか..！？

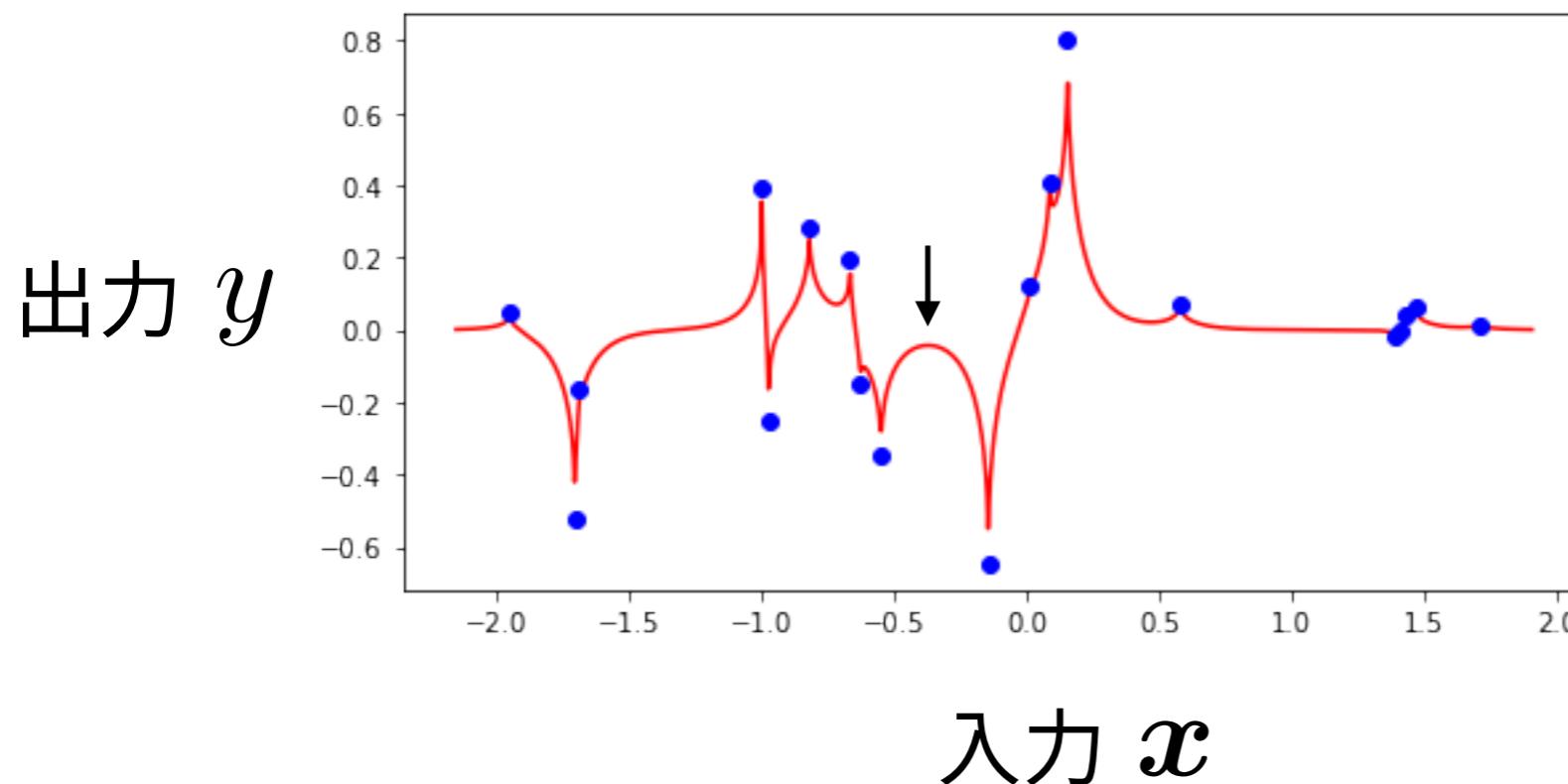
アーティファクト？

# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



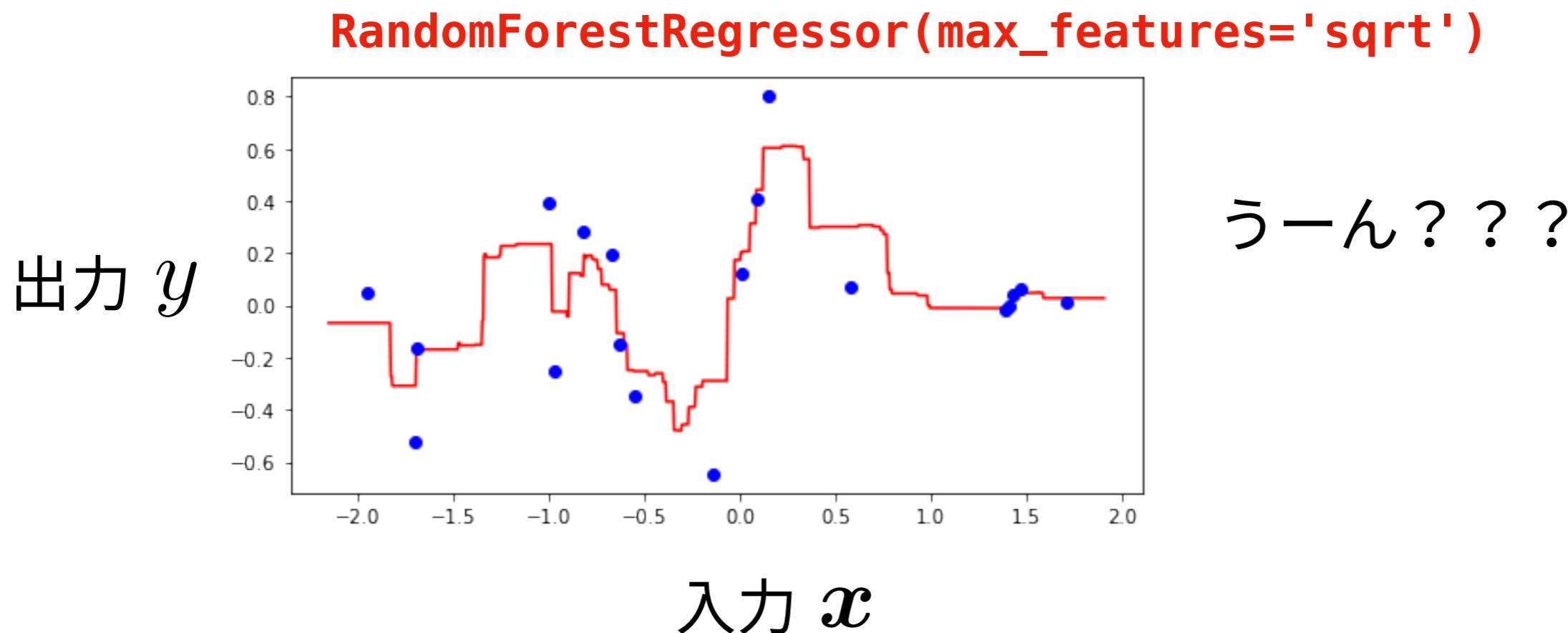
`GaussianProcessRegressor(kernel=Matern(length_scale=100.0, nu=0.2))`



うーん？？？  
特に ↓ の箇所は  
これで良いのか..！？  
アーティファクト？

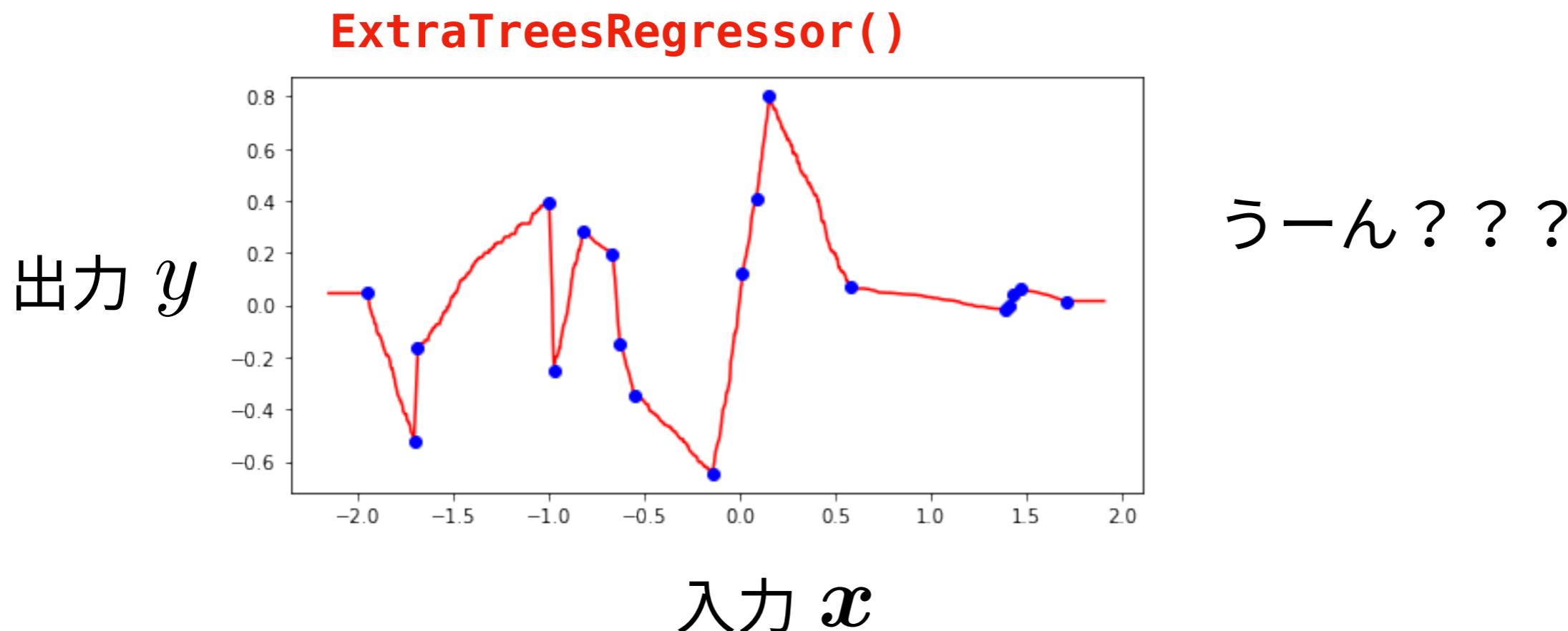
# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力  $x$

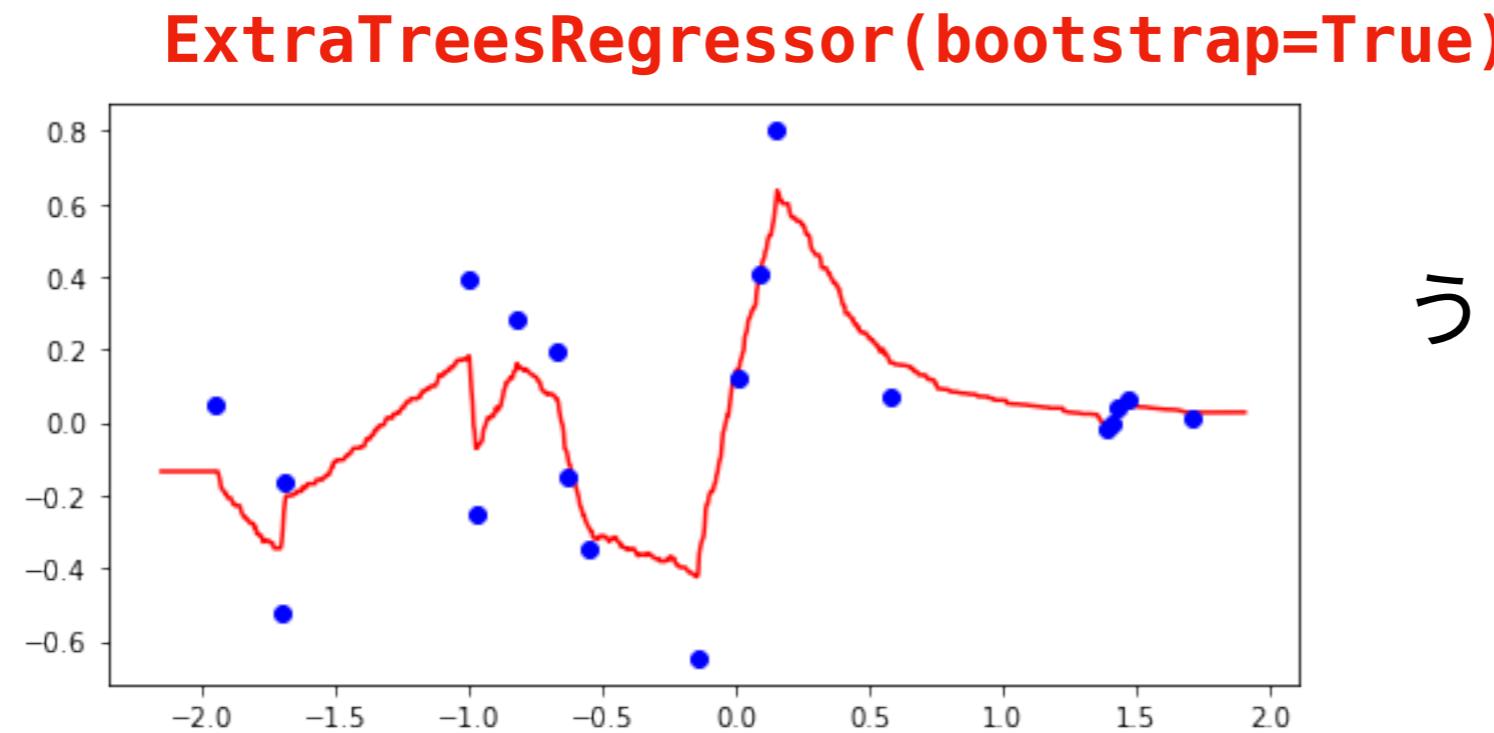
- 元素組成比
- 実験条件

機械学習モデル

出力  $y$

- 収率

出力  $y$



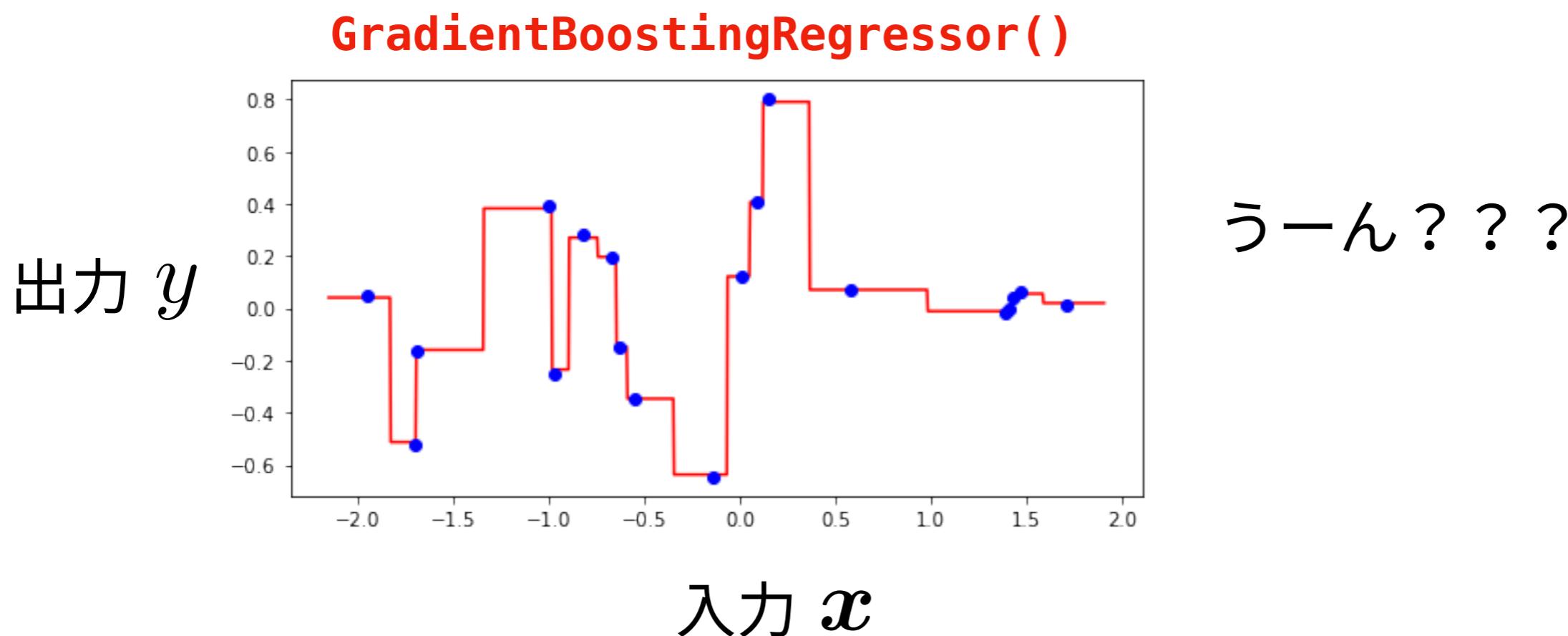
入力  $x$

sklearnのデフォルト  
はoffなので注意

うーん？？？

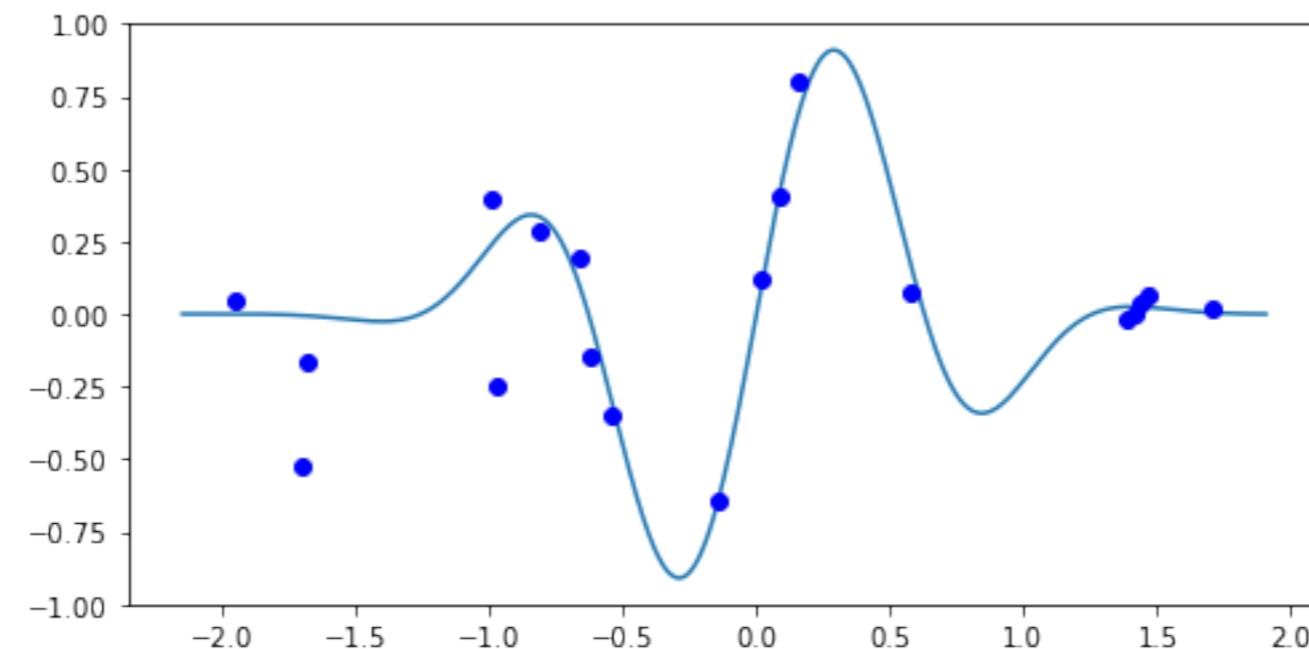
# このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



この例は実は「真のモデル+ノイズ」の人工データ

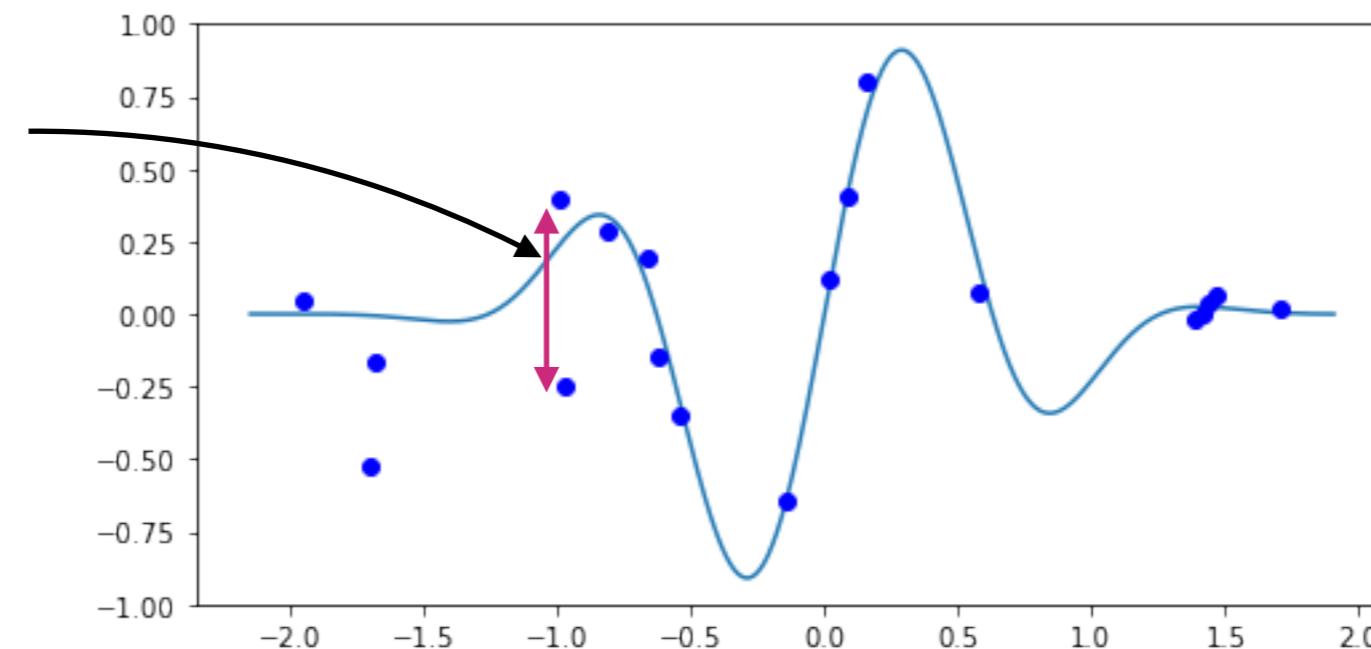
現実の実測データ(ましてや報告データ)には色々な落とし穴が！



# この例は実は「真のモデル+ノイズ」の人工データ

現実の実測データ(ましてや報告データ)には色々な落とし穴が！

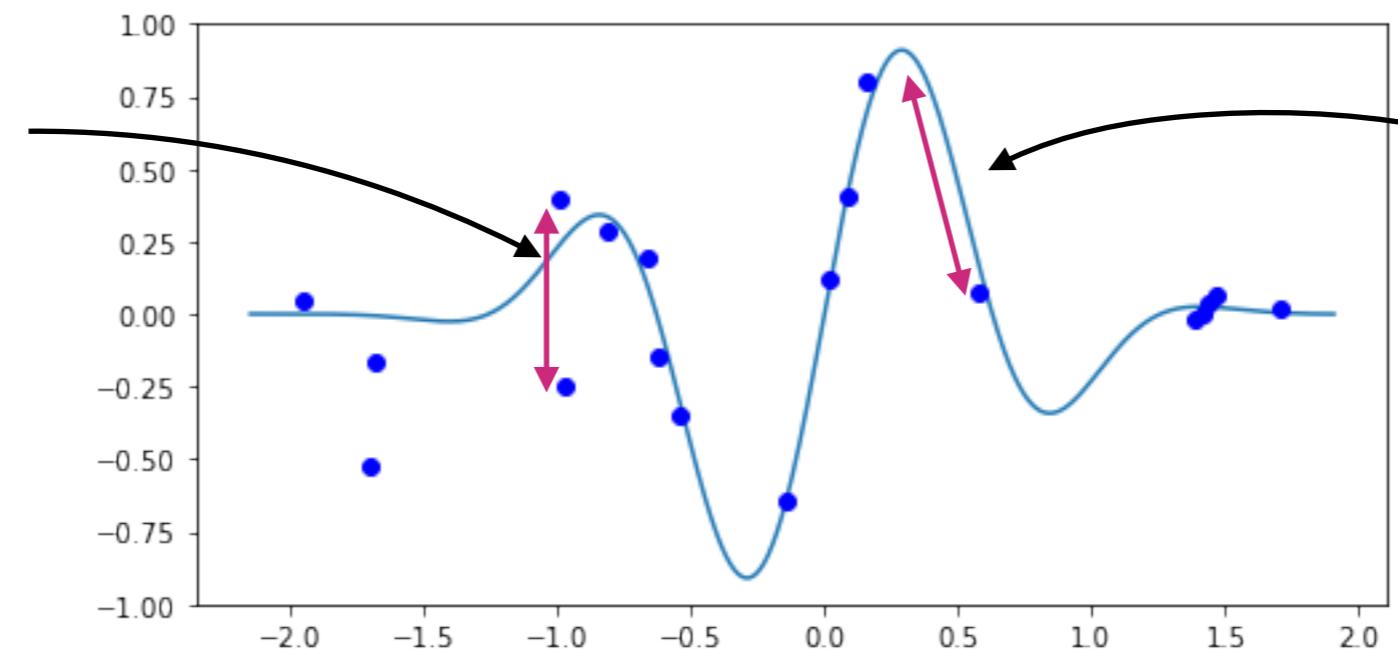
近い正解例が  
inconsistent  
(教師ノイズ)



# この例は実は「真のモデル+ノイズ」の人工データ

現実の実測データ(ましてや報告データ)には色々な落とし穴が！

近い正解例が  
inconsistent  
(教師ノイズ)

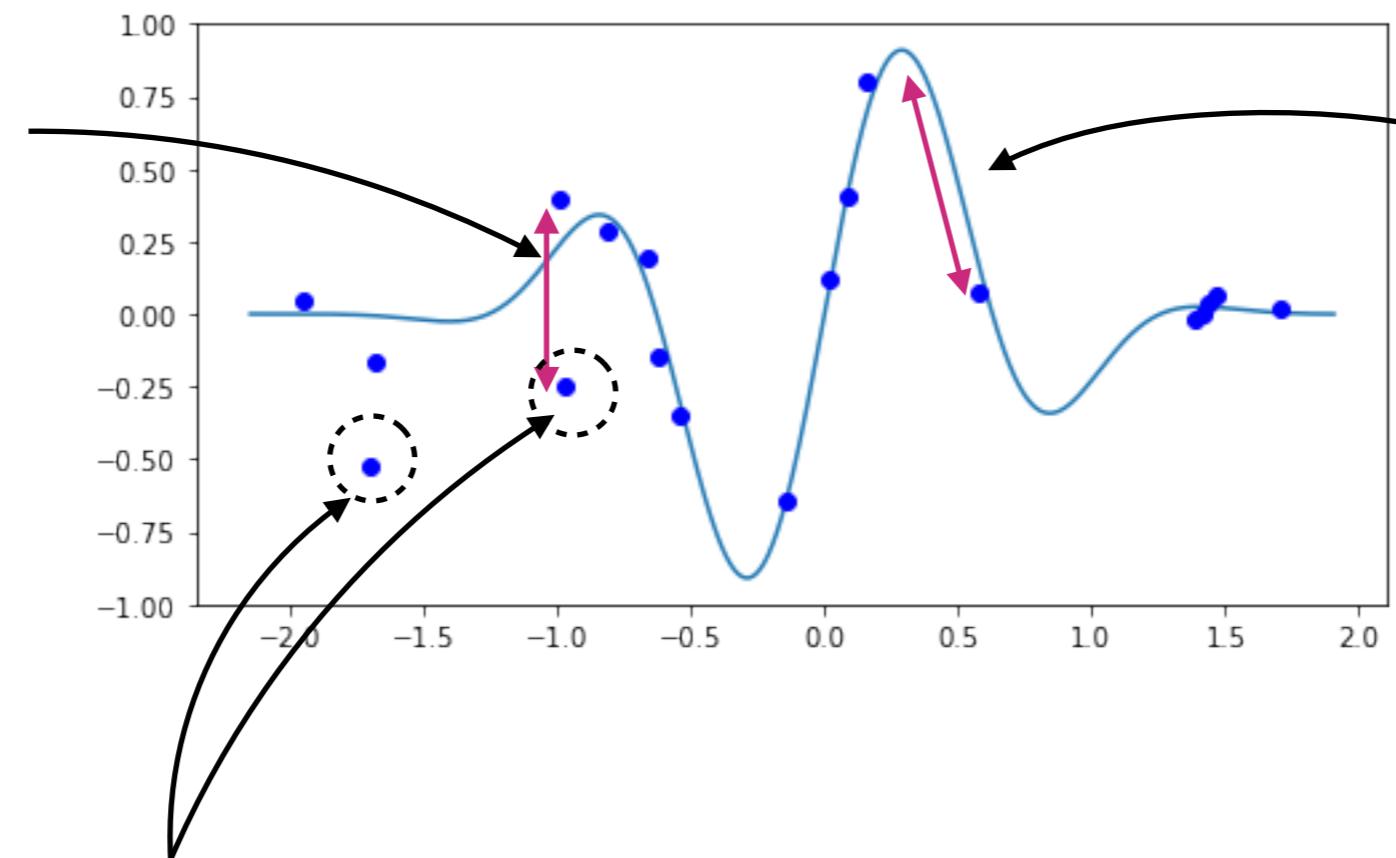


xの少しの違いで  
yに急峻な変化  
(Cliffs)

# この例は実は「真のモデル+ノイズ」の人工データ

現実の実測データ(ましてや報告データ)には色々な落とし穴が！

近い正解例が  
inconsistent  
(教師ノイズ)



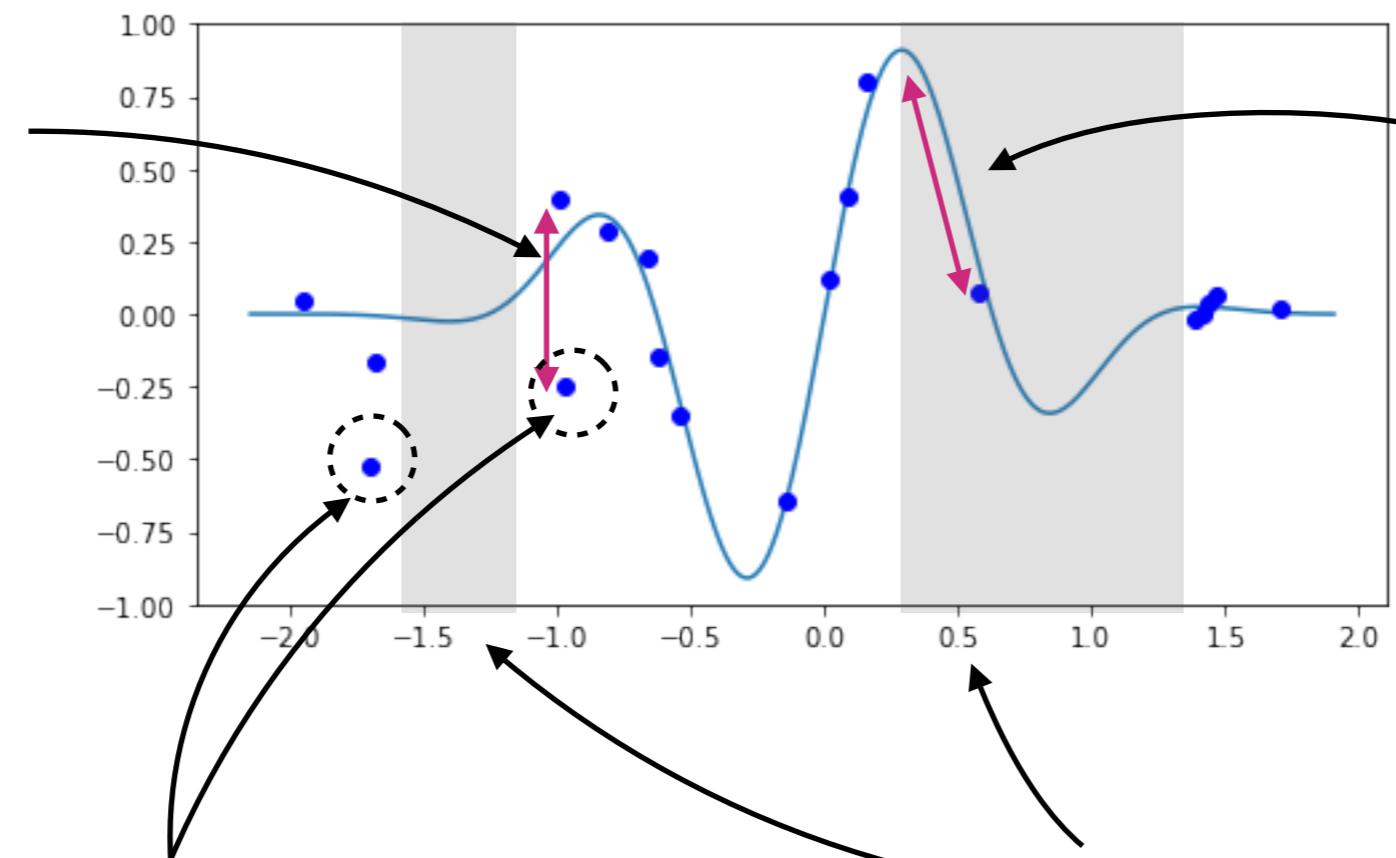
外れ値

xの少しの違いで  
yに急峻な変化  
(Cliffs)

# この例は実は「真のモデル+ノイズ」の人工データ

現実の実測データ(ましてや報告データ)には色々な落とし穴が！

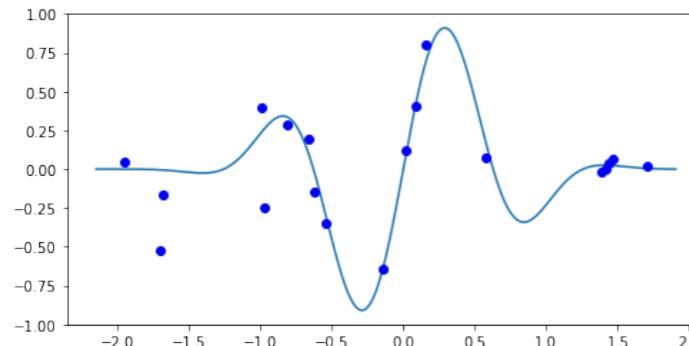
近い正解例が  
inconsistent  
(教師ノイズ)



xの少しの違いで  
yに急峻な変化  
(Cliffs)

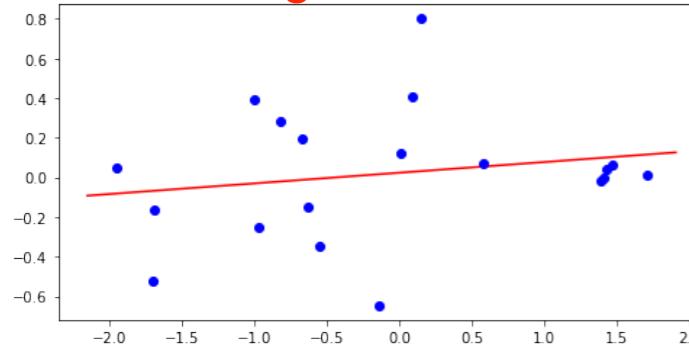
データがない/少ない領域

# この例は実は「真のモデル+ノイズ」の人工データ

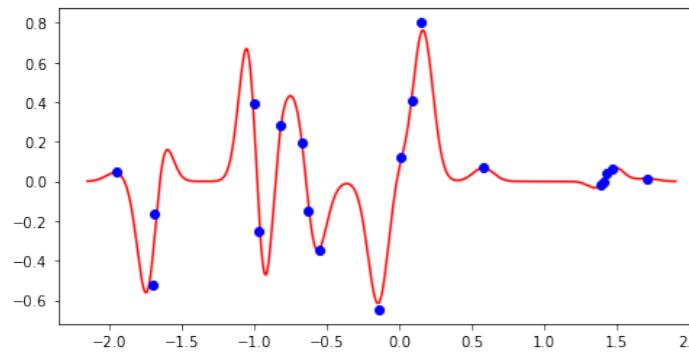


← “真の”モデル

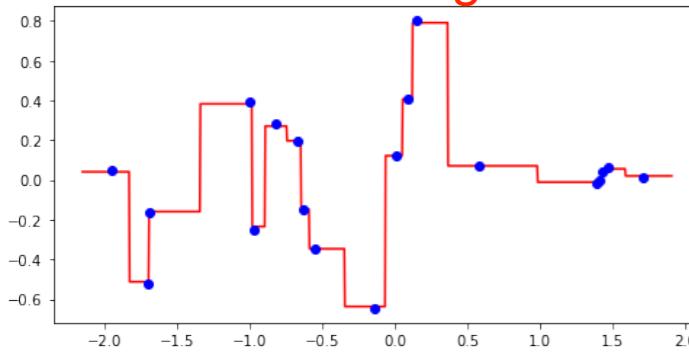
Linear Regression



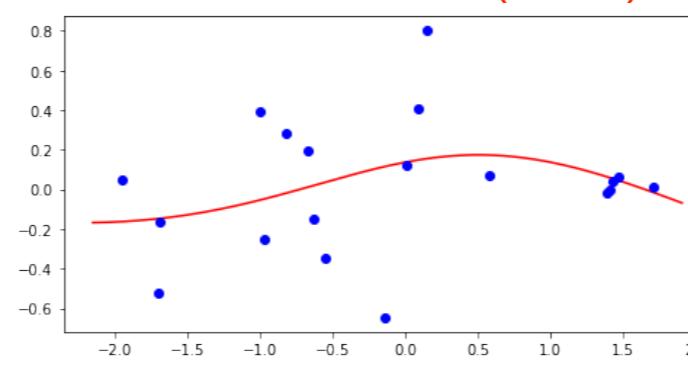
Kernel Ridge (RBF)



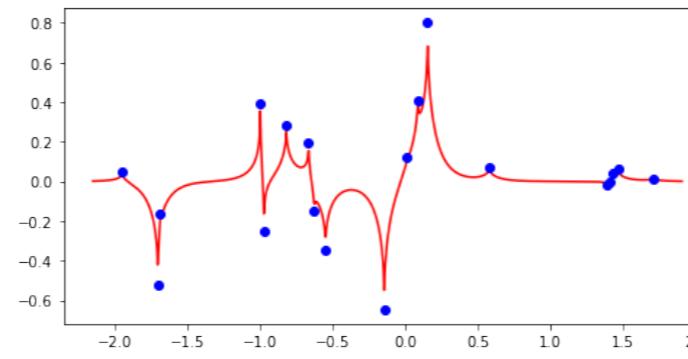
Gradient Boosting



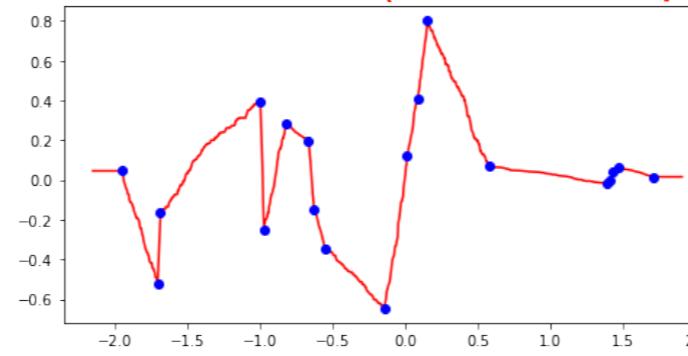
Neural Networks (Tanh)



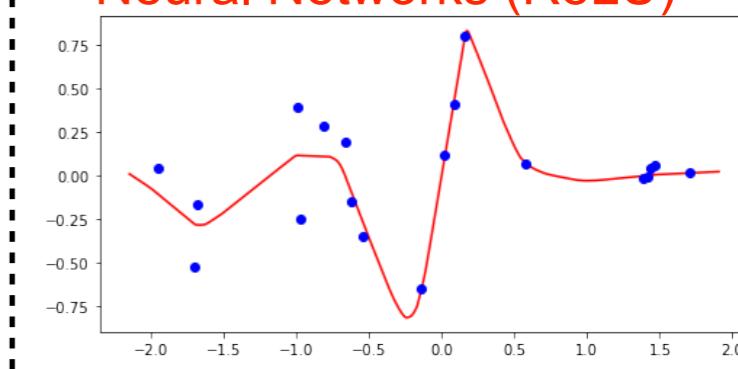
Kernel Ridge (Laplacian)



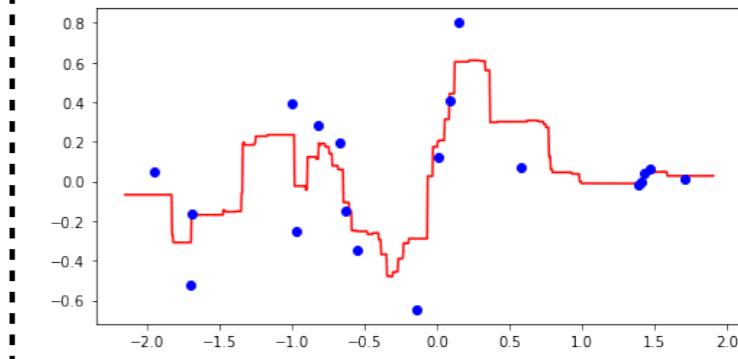
Extra Trees (no bootstrap)



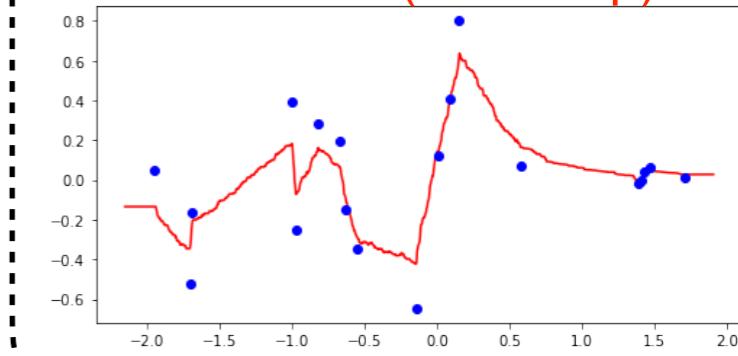
Neural Networks (ReLU)



Random Forest



Extra Trees (bootstrap)



# Rashomon効果：一体どれを信じればいいんじやい！

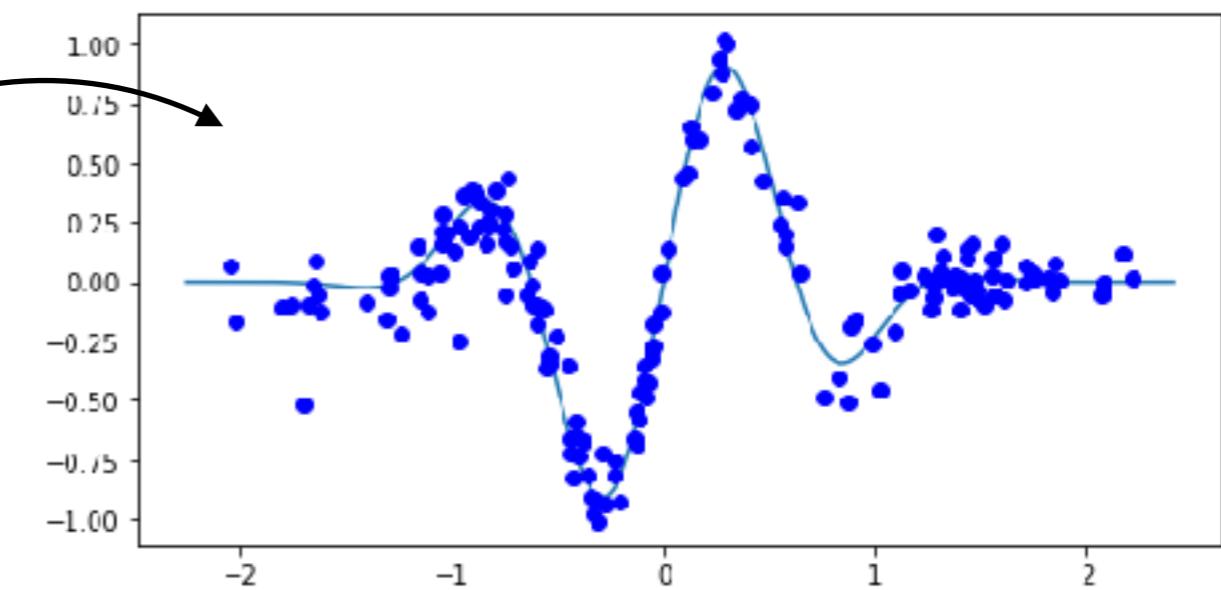
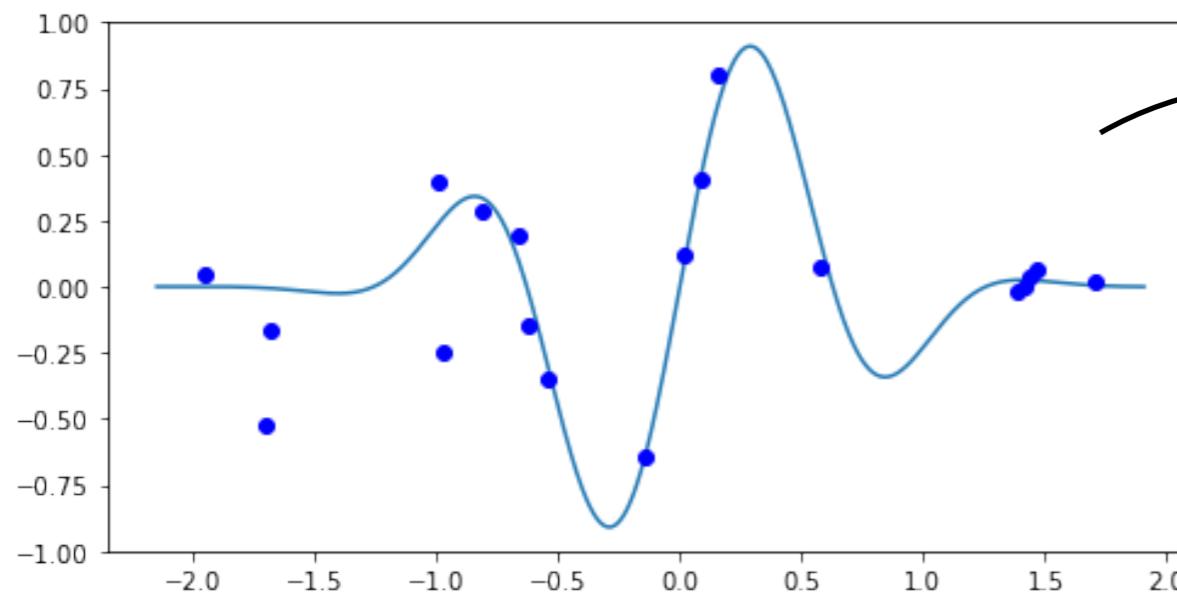
🤔 **機械学習モデルや訓練データが変われば予測は変わる**  
(モデルとデータの数だけ予測もある = 真実は「數の中」)

- Cross-validation精度はほぼ同等のモデルが無数にあり得る
- 同じモデルでもHyperparameterが違えばかなり違い得る
- 現実では真のモデルは分からぬため良し悪しの判断は困難
- 入力変数を思いつくだけ入れて高次元データになるとさらにRashomon効果リスクが増大 (私の1次元例の作為性にご注意!)

MLがどういう技術なのか**MLの特性と限界**を正しく把握する  
予測結果を当該分野の知識に照らして注意深く検証・解釈する

# ちなみにサンプル数が十分大きければわりとどれでもOK

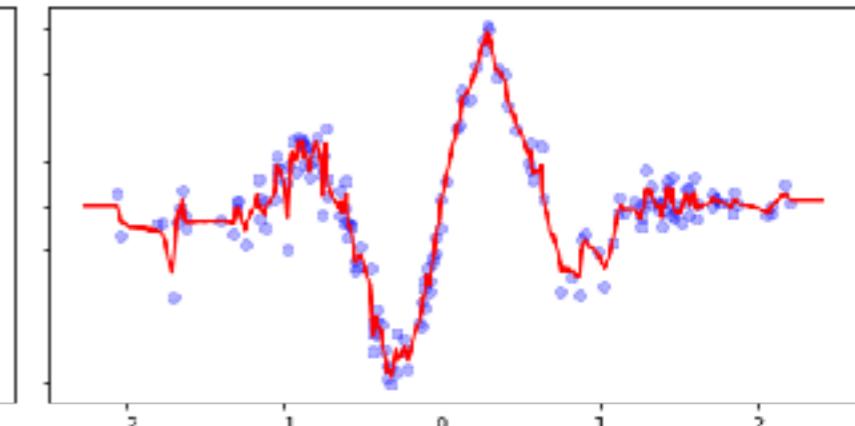
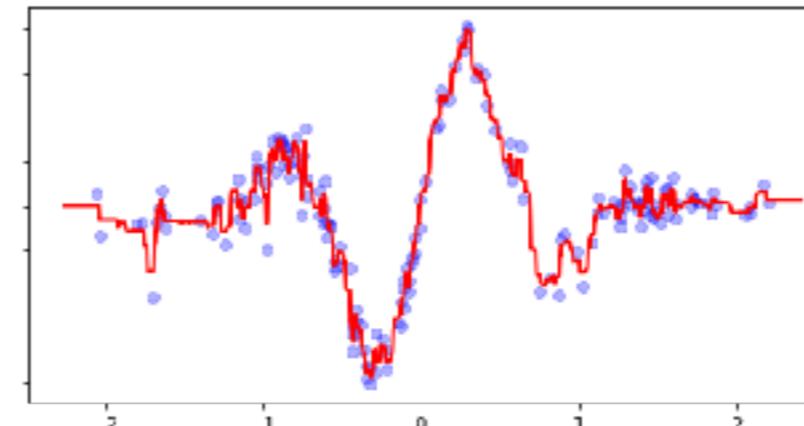
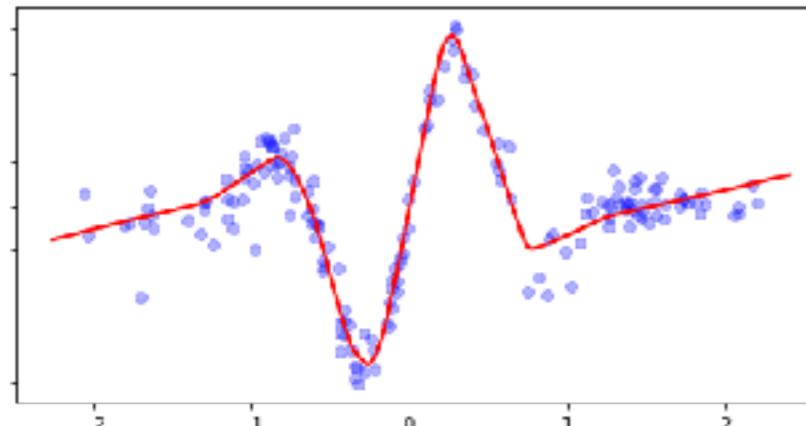
**サンプル数**が十分多ければ例外的データ点は統計的に相殺される  
 → ただし高次元であるほど**指数的な数が必要**で非現実的な期待



Neural Networks (ReLU)

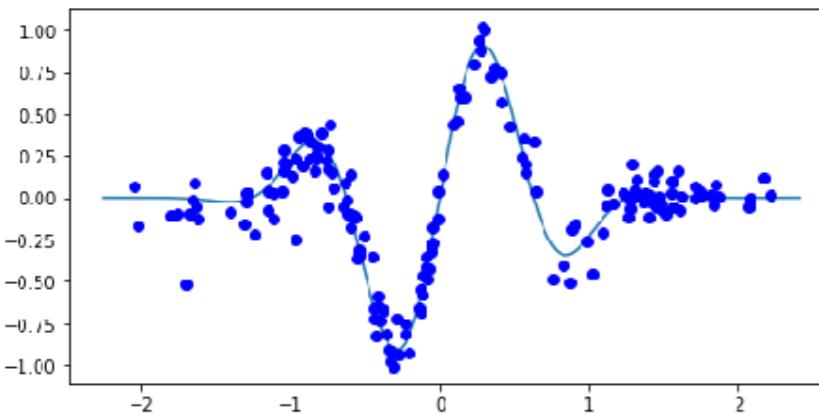
Random Forest

Extra Trees (bootstrap)

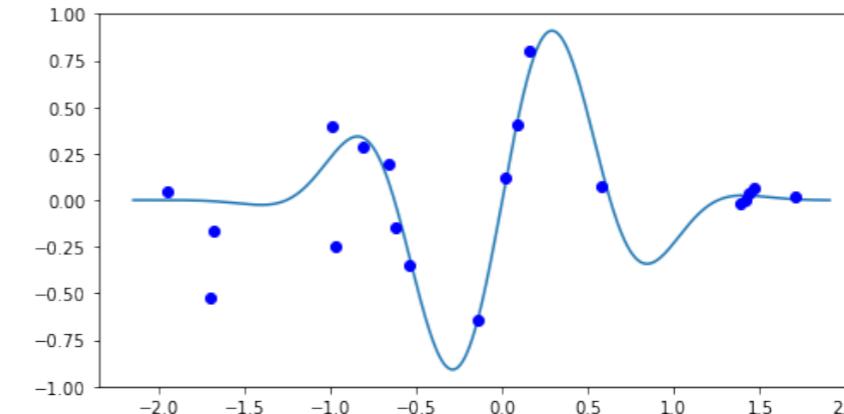


# 多くは念頭にある候補空間に対しデータが足りていない！

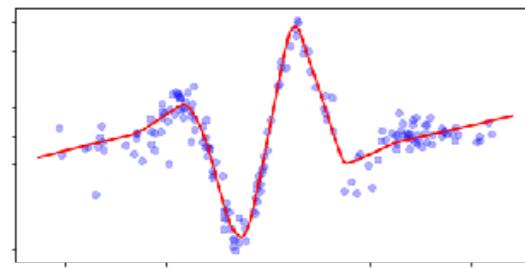
理想的なデータ



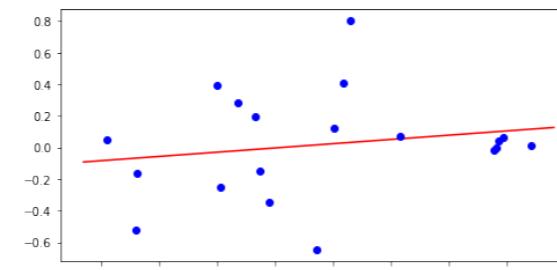
実際に手に入るデータ (Underspecification)



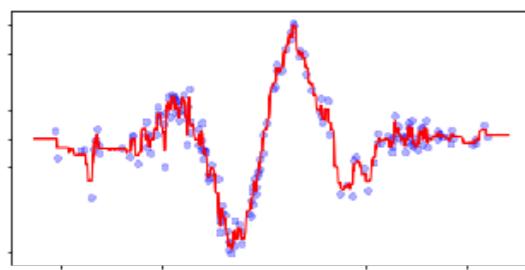
Neural Networks (ReLU)



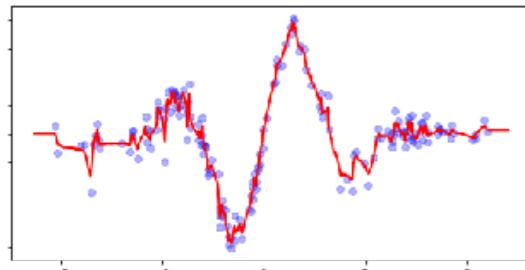
Linear Regression



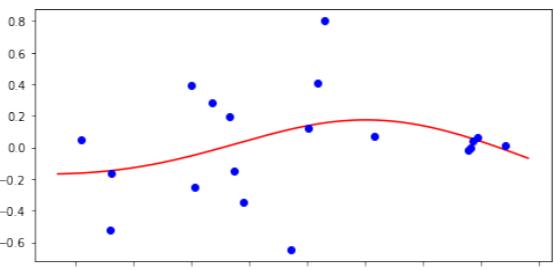
Random Forest



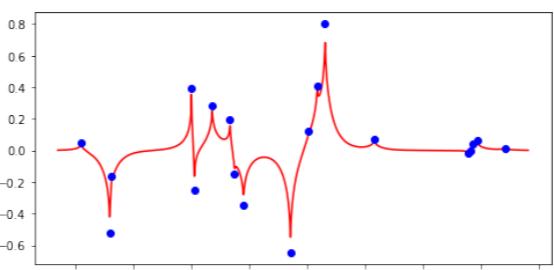
Extra Trees (bootstrap)



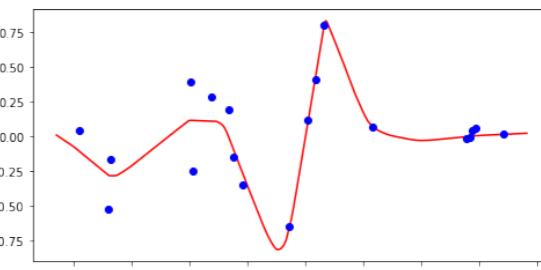
Neural Networks (Tanh)



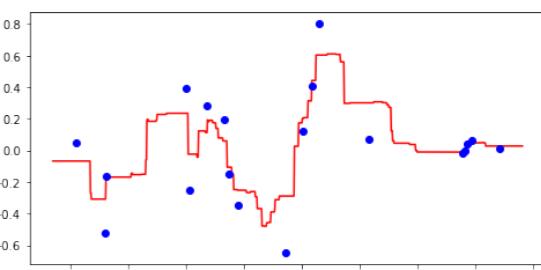
Kernel Ridge (Laplacian)



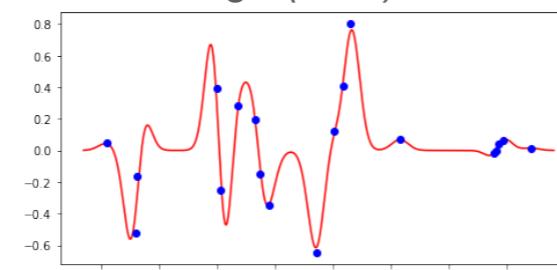
Neural Networks (ReLU)



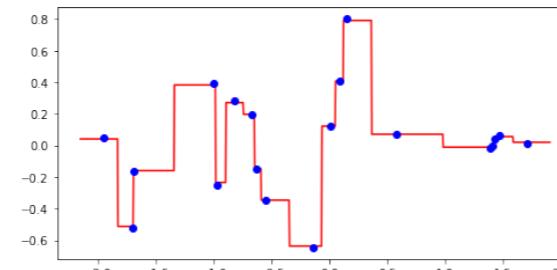
Random Forest



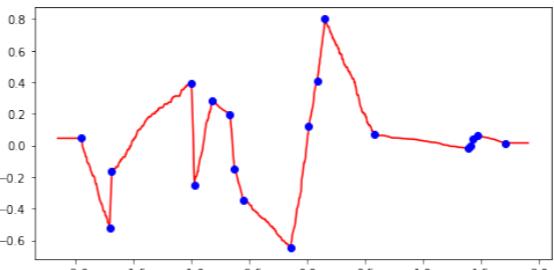
Kernel Ridge (RBF)



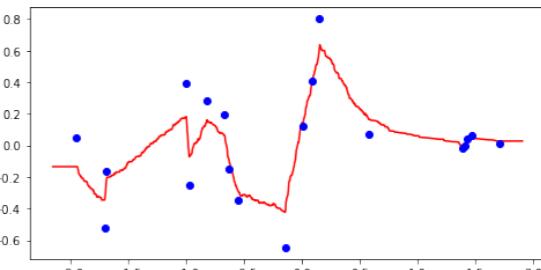
Gradient Boosting



Extra Trees (no bootstrap)

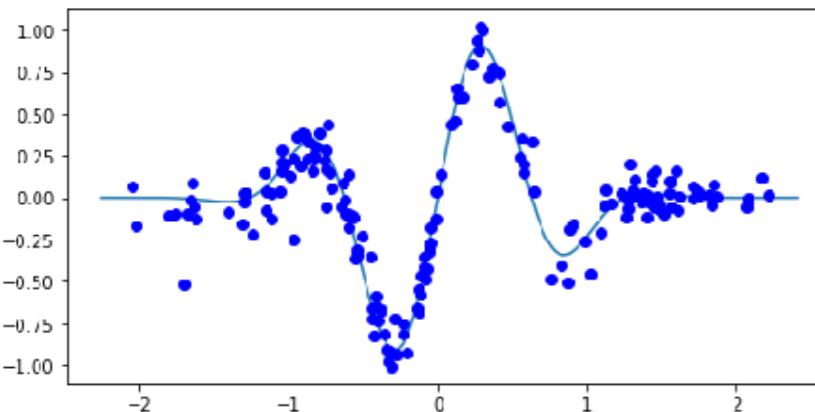


Extra Trees (bootstrap)

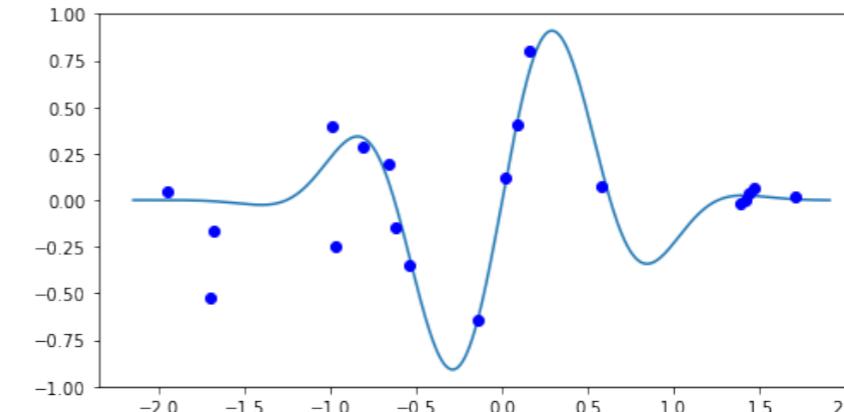


# 多くは念頭にある候補空間に対しデータが足りていない！

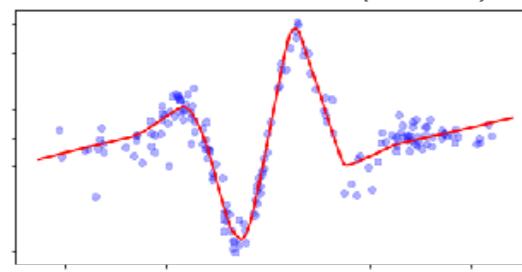
理想的なデータ



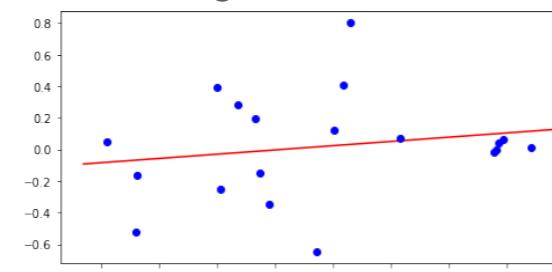
実際に手に入るデータ (Underspecification)



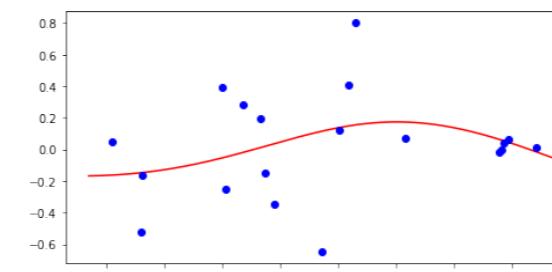
Neural Networks (ReLU)



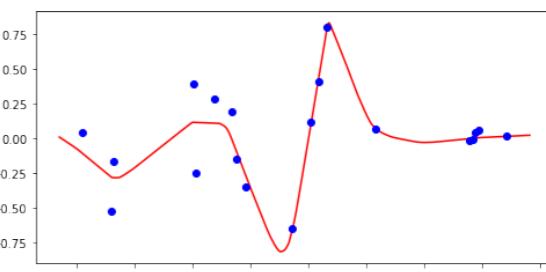
Linear Regression



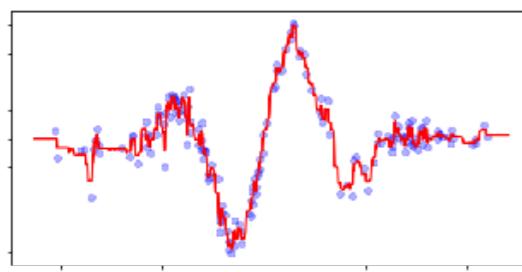
Neural Networks (Tanh)



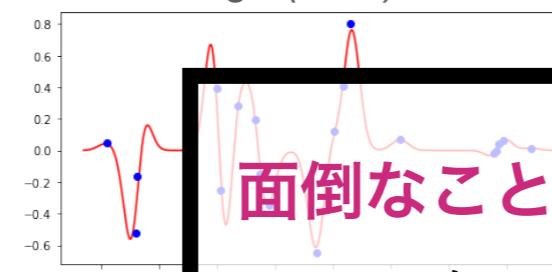
Neural Networks (ReLU)



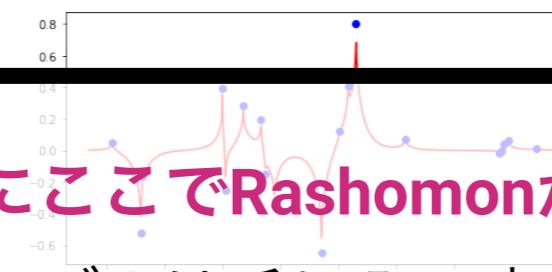
Random Forest



Kernel Ridge (RBF)



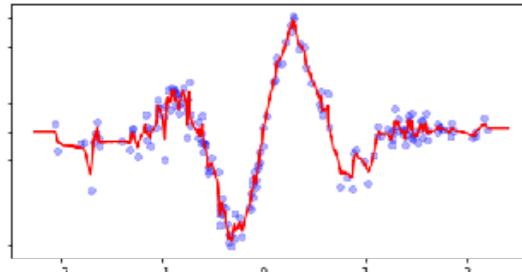
Kernel Ridge (Laplacian)



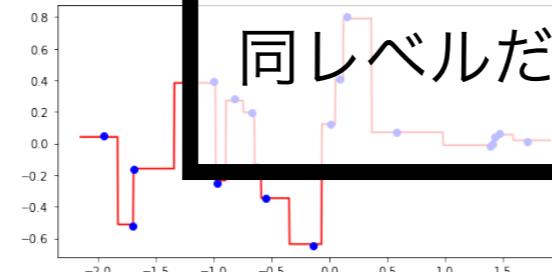
Random Forest



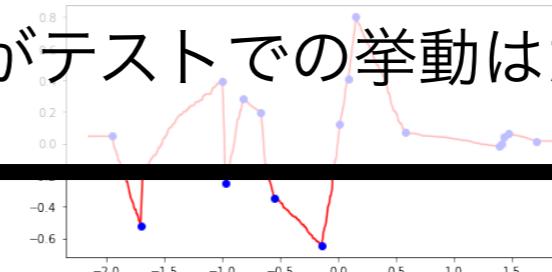
Extra Trees (bootstrap)



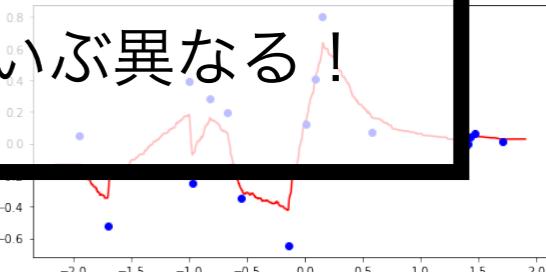
Gradient Boosting



Extra Trees (no bootstrap)



Extra Trees (bootstrap)



面倒なことにここでRashomonが起きてしまう！

このへんのモデルは手に入る事例でのCV精度は  
同レベルだがテストでの挙動はだいぶ異なる！

# Leo Breimanの3レッスン

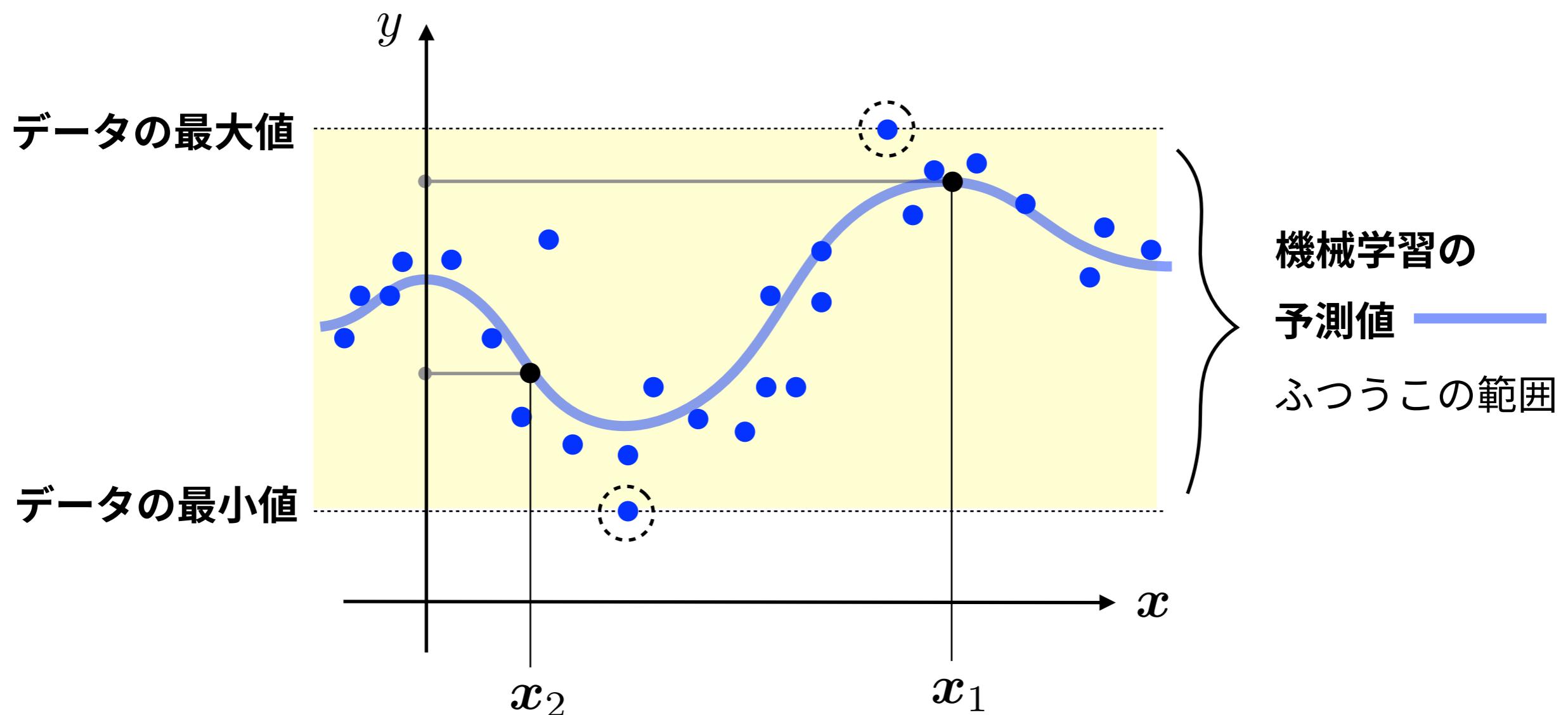
## Rashomon, Occam, Bellman

今でもこの3点はまだ色々とOpenな問題を孕んで研究されている

- "Rashomon": 良いモデルの多重性(非一意性)  
同程度の良い予測精度を持つ全く異なるモデルがたくさん存在する
- "Occam": モデルの解釈性と予測精度のコンフリクト  
モデルのシンプルさ(解釈性)と予測精度の両立はとても難しい
- "Bellman": 高次元データが引き起こすメリットとデメリット  
高次元な表現(関係しそうなできるだけ多くの変量)を扱うべきなのか  
伝統的な統計学のように支配的な少数の変量を検討し分析すべきなのか  
→ キッチングシンク回帰(思いつく変数全部入りモデル)・特徴量  
エンジニアリングと擬似相関・Rashomon効果の増大リスク

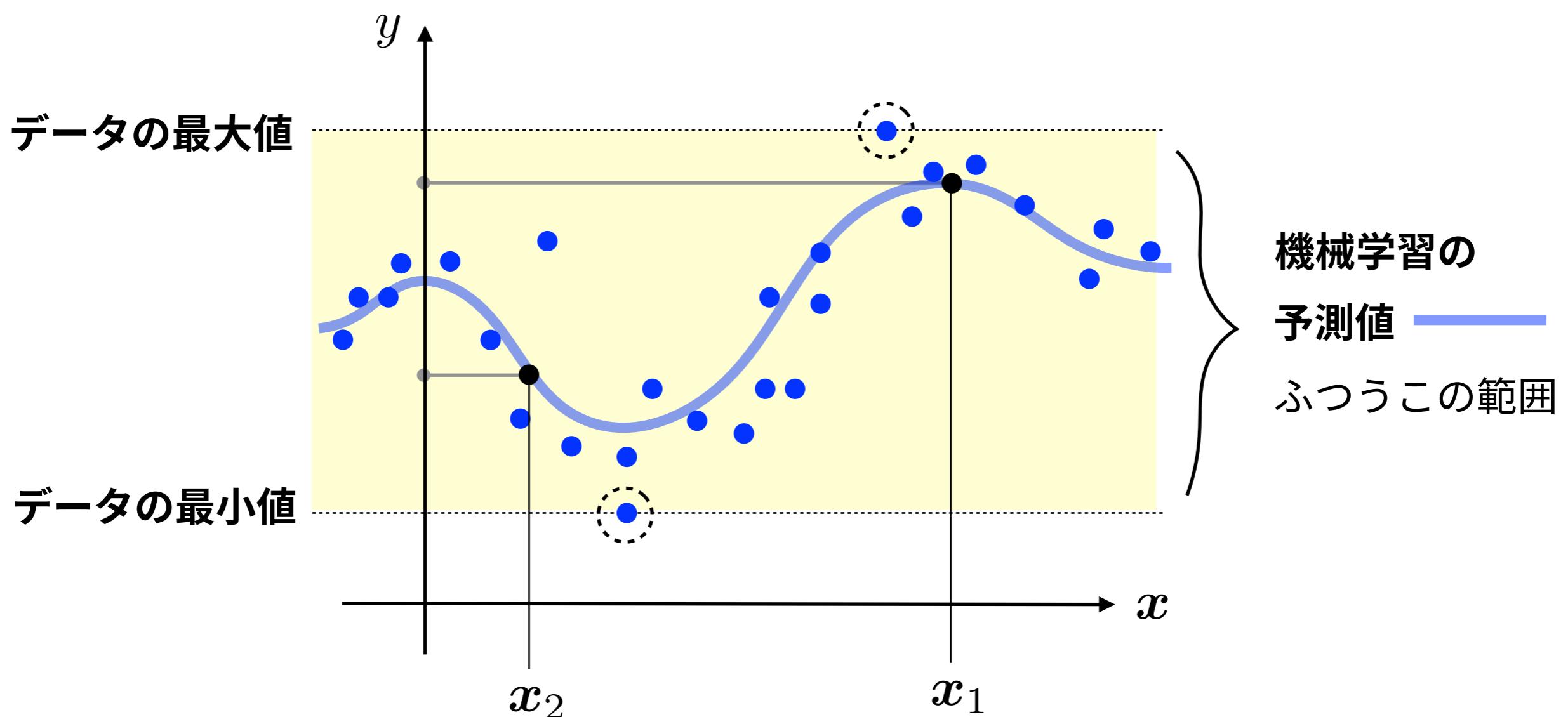
# 機械学習の予測値は訓練データの最良値をふつう超えない

機械学習モデルは期待誤差が最小になるよう(訓練データの真ん中を通るよう)にフィッティングされるため



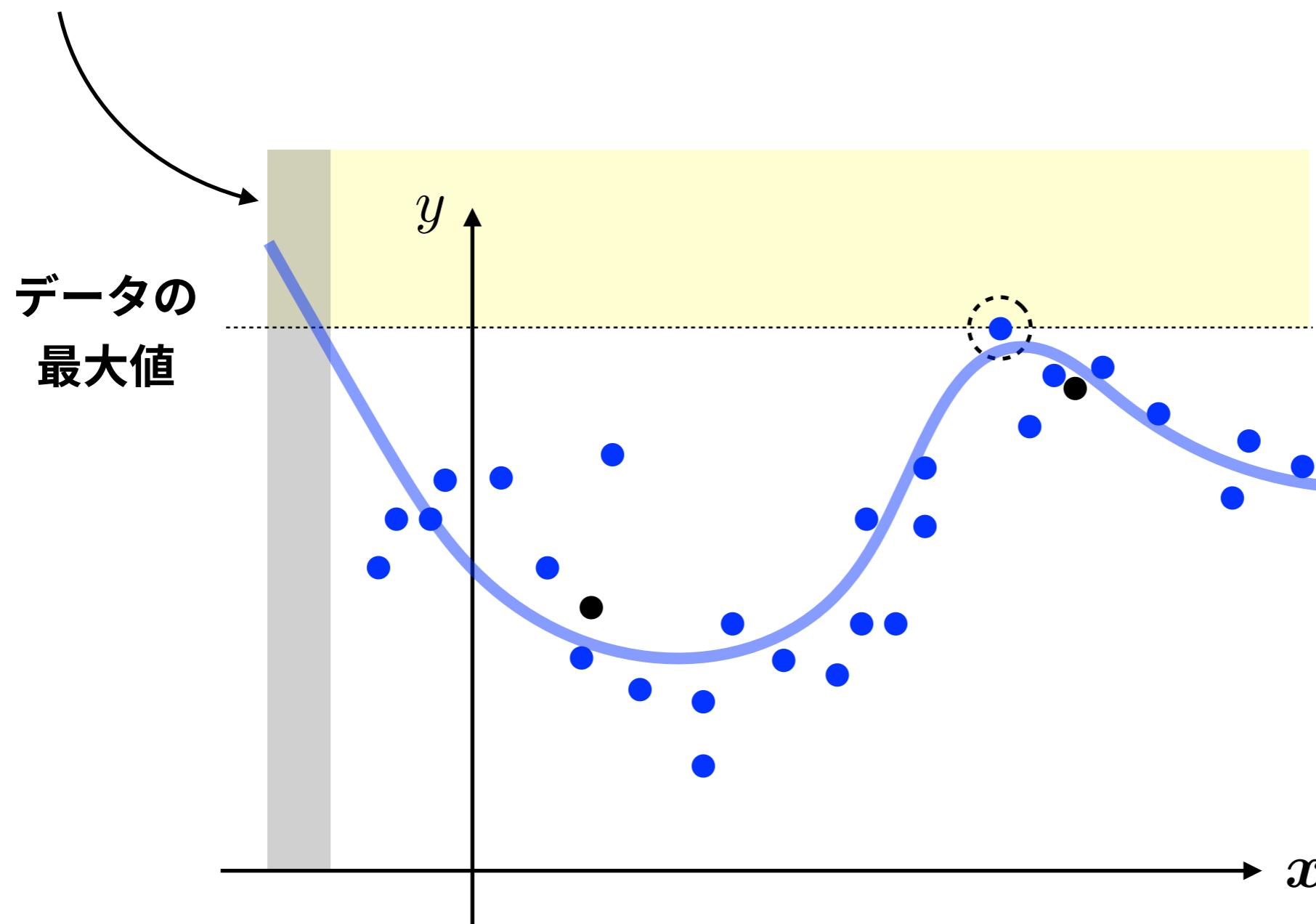
# 機械学習の予測値は訓練データの最良値をふつう超えない

機械学習モデルは期待誤差が最小になるよう<sup>(訓練データの真ん中を通るよう)</sup>にフィッティングされるため  
→ 既知のものより良いものを見つけるという探索目的に不適合



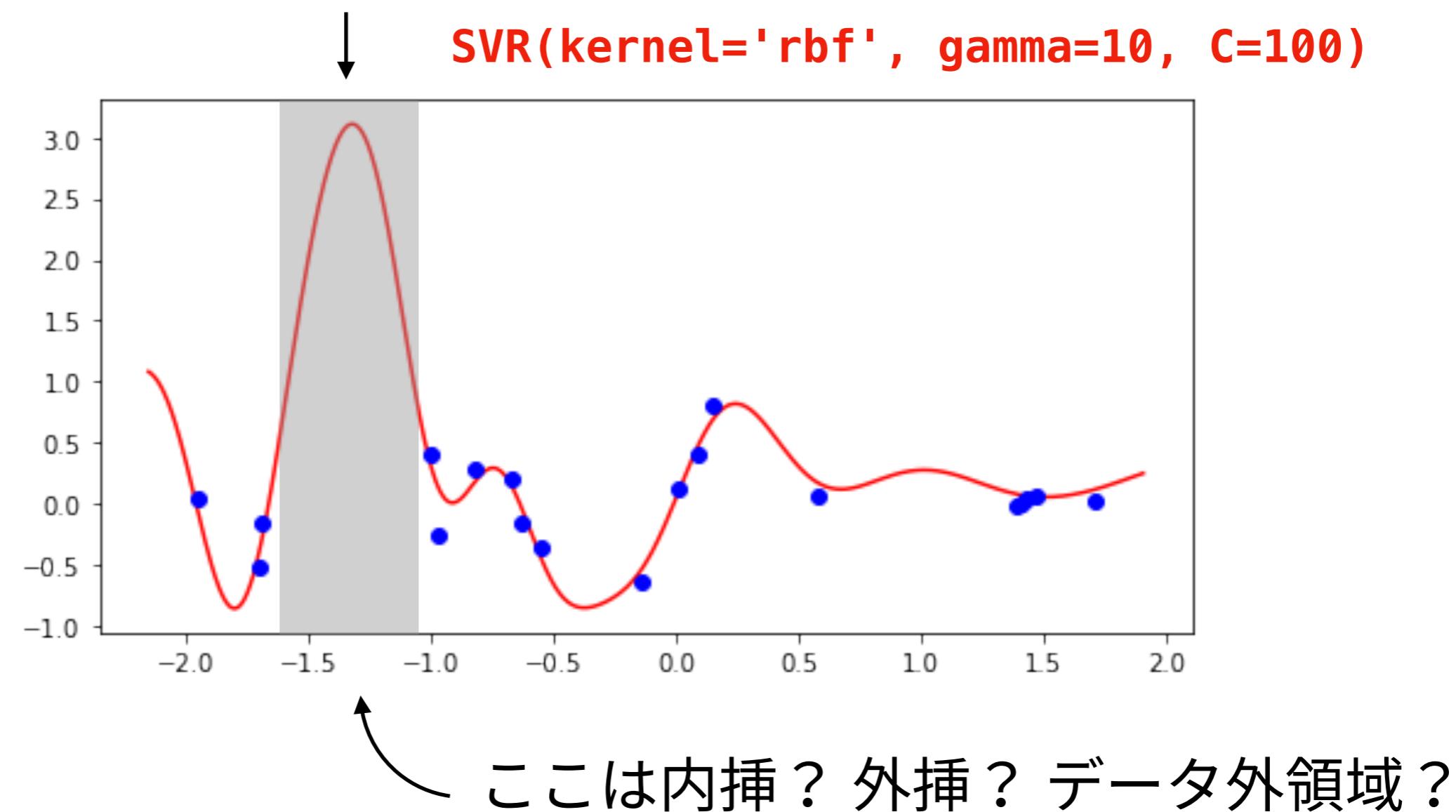
# 機械学習の予測値は訓練データの最良値をふつう超えない

さらに、もし予測値が訓練データの最良値を上回るとしても、データがない領域でその予測は任意的で当てにできない…



# 機械学習の予測値は訓練データの最良値をふつう超えない

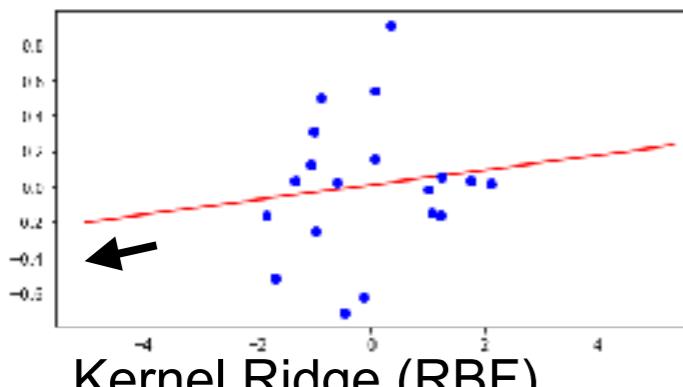
1次元の例では「内挿」か「外挿」か分かる感じがしてしまうが  
一般的の高次元ではこの判別すら直感的ではないことに注意



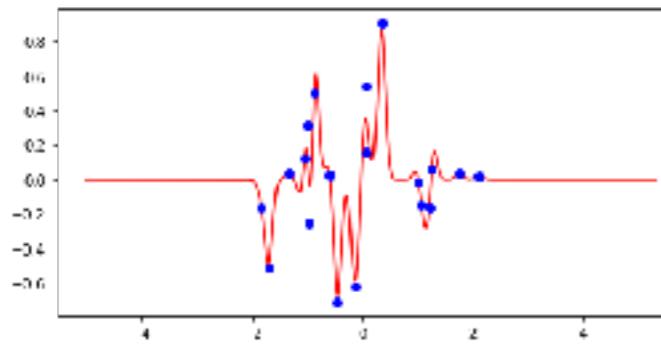
# 「データ外領域での意図しない外挿」リスクは手法に依存

決定木アンサンブル法(Random Forest等)ではこの状況は原理上起こらないが、線形回帰やNeural Networksなど他の手法では注意

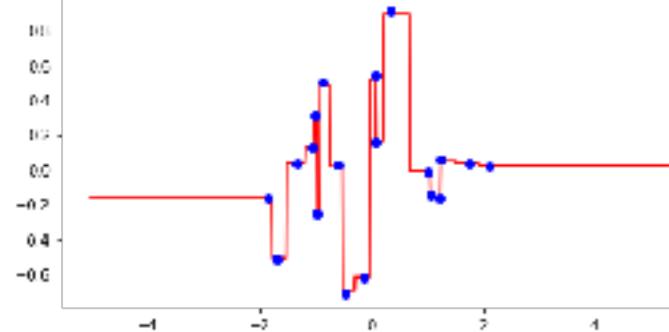
Linear Regression



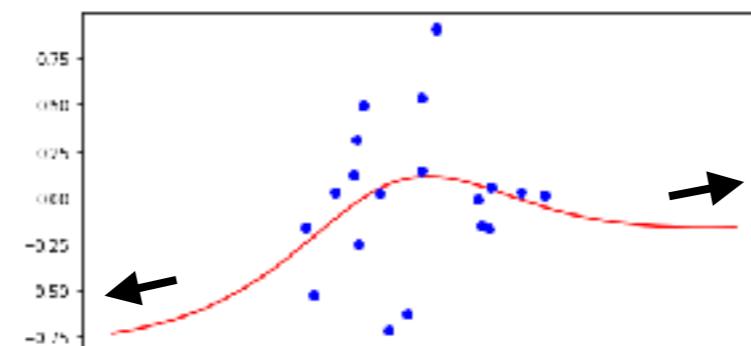
Kernel Ridge (RBF)



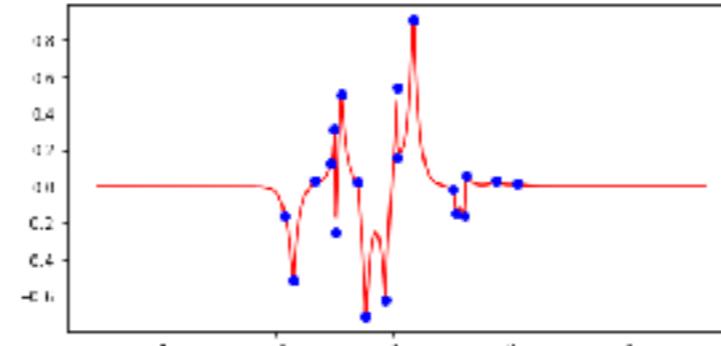
Gradient Boosting



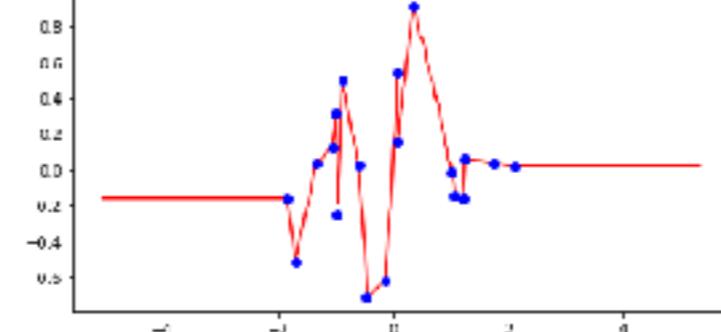
Neural Networks (Tanh)



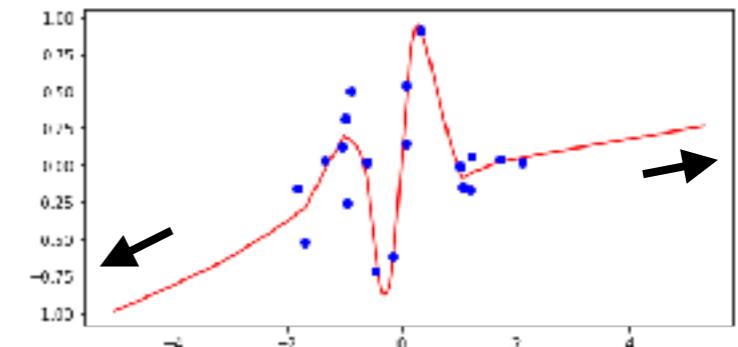
Kernel Ridge (Laplacian)



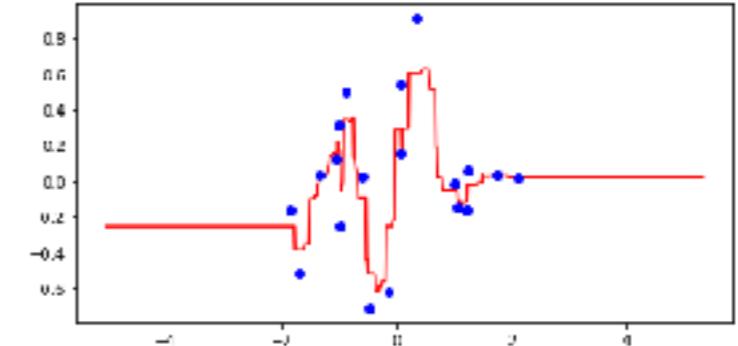
Extra Trees (no bootstrap)



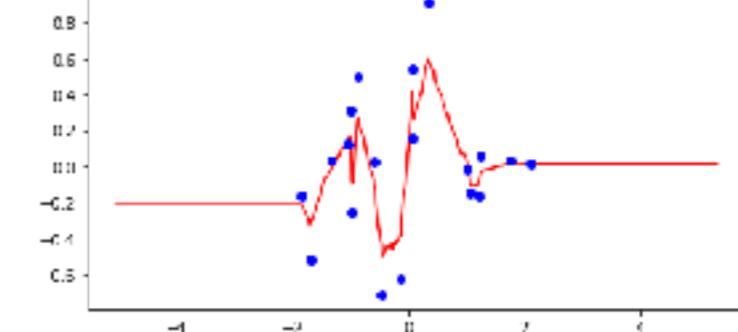
Neural Networks (ReLU)



Random Forest



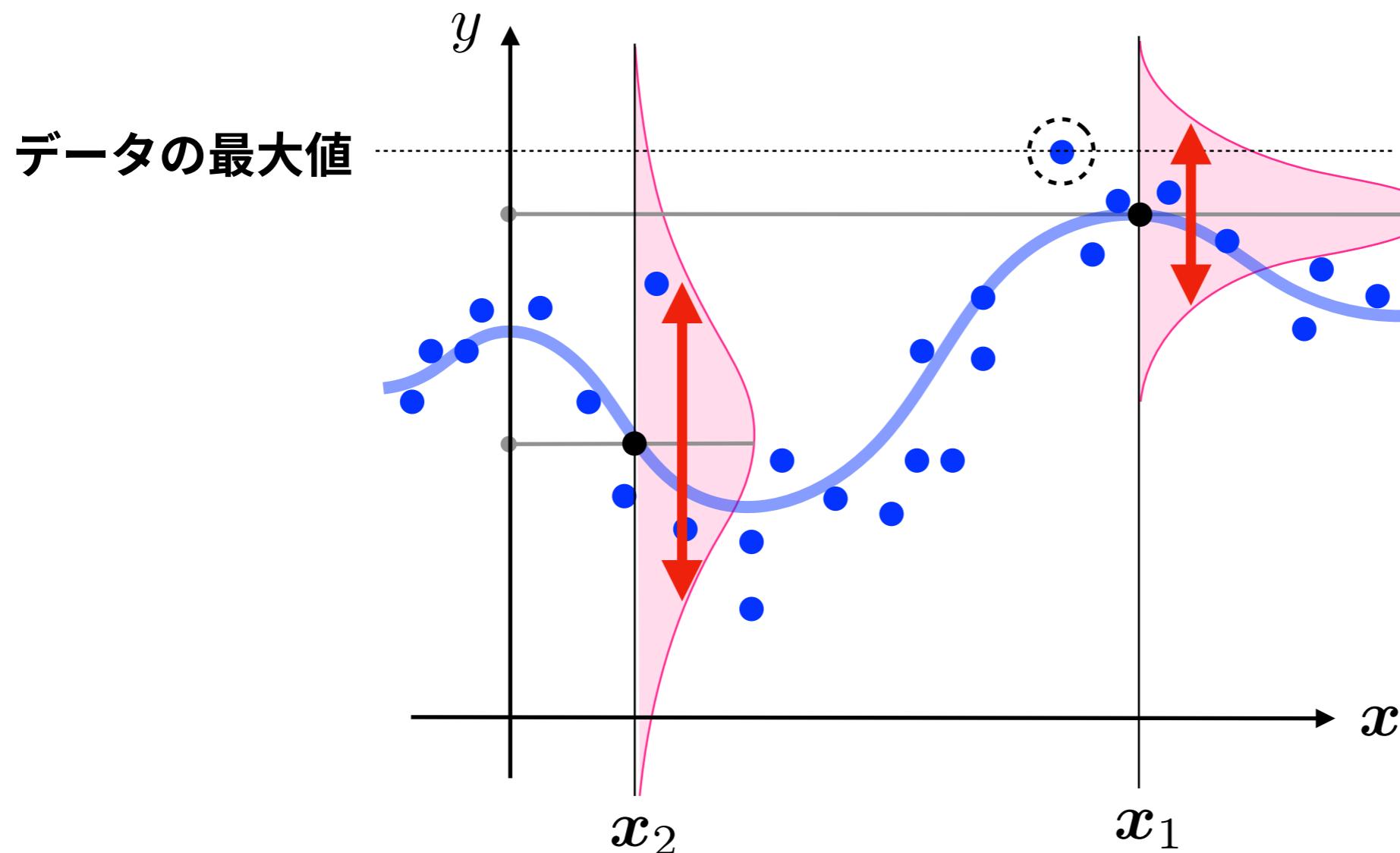
Extra Trees (bootstrap)



# 予測値だけではなくその予測分散(確度)も考えるのが重要

探索に関する意思決定に活用するのであれば機械学習モデルの  
予測値の分散/分布/信頼区間を考えることが重要

e.g. 「収率予測値は  $20.2 \pm 15.1$ 」 vs 「収率予測値は  $20.2 \pm 2.5$ 」

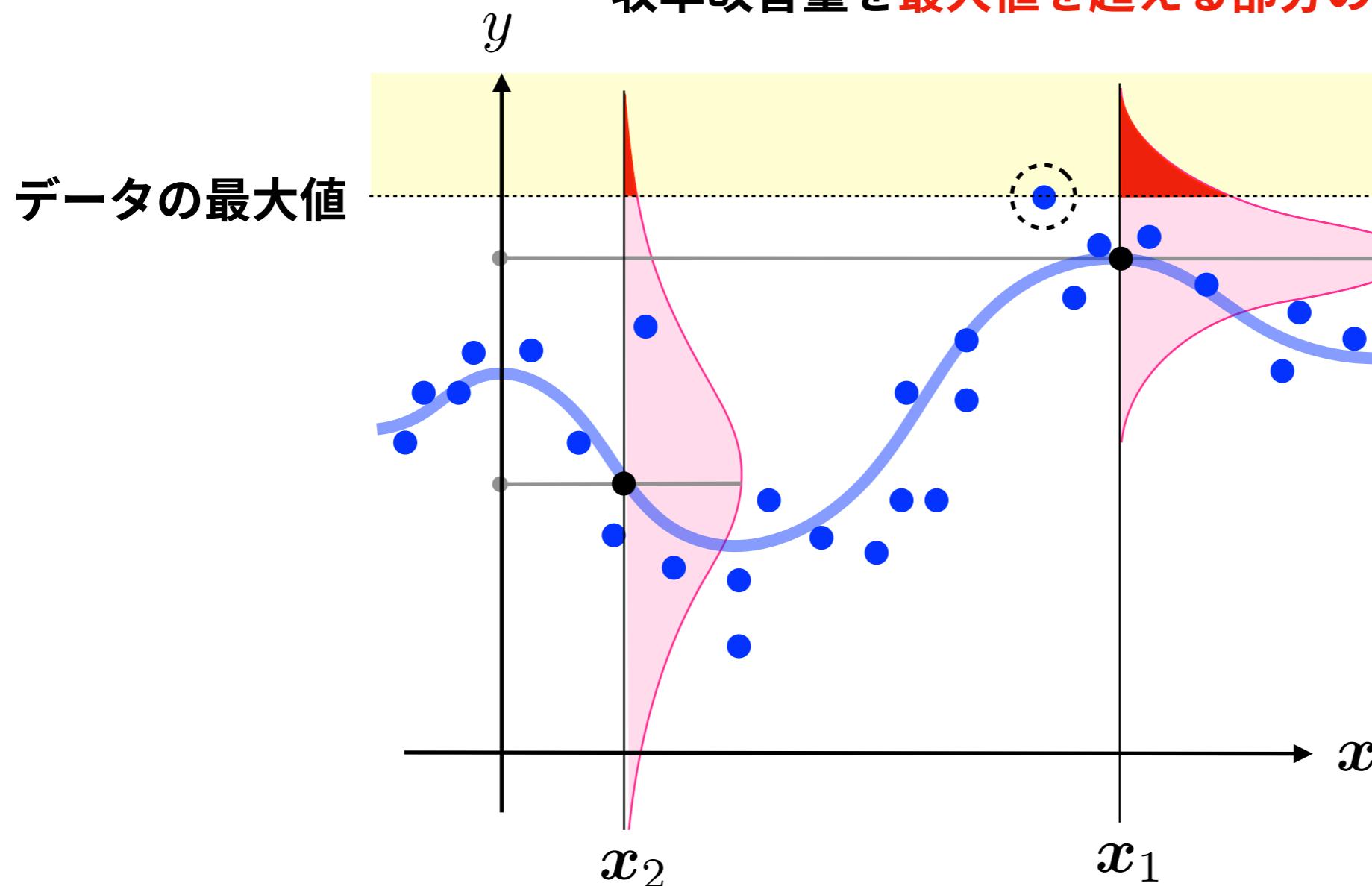


# 探索する際は予測値 자체を指標としない

探索の目的では期待改善(EI)や信頼区間の上限などを指標に

期待改善(EI) = 収率改善量の期待値

収率改善量を最大値を超える部分の確率の重みで積分



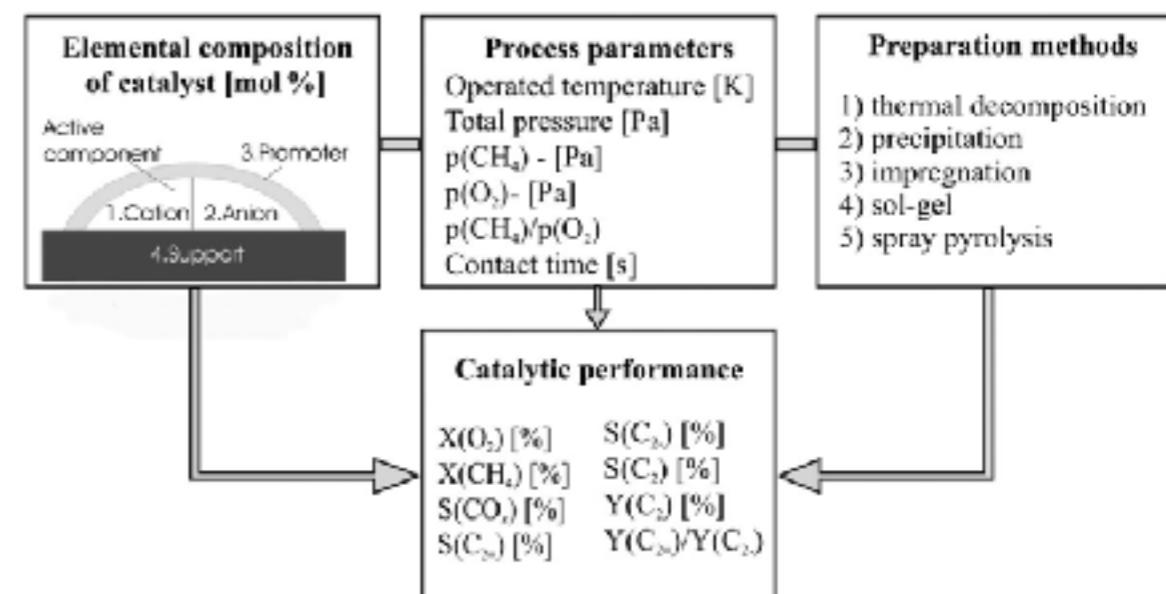
# 文献から集めた実際の実験データ報告を使う

# 1866 catalyst records from 421 reports

# Oxidative coupling of methane (OCM) reactions

**Methane ( $\text{CH}_4$ )** is partially oxidized to  **$\text{C}_2$  hydrocarbons** such as ethane ( $\text{C}_2\text{H}_6$ ) and ethylene ( $\text{C}_2\text{H}_4$ ) in a single step

- Zavyalova, U.; Holena, M.; Schlögl, R.; Baerns, *ChemCatChem* 2011.
  - Followup:  
Kondratenko, E. V.; Schlüter, M.; Baerns, M.; Linke, D.; Holena, M.  
*Catal. Sci. Technol.* 2015.
  - Renalysis with Corrections & Outlier Removal  
Schmack, R.; Friedrich, A.; Kondratenko, E. V.; Polte, J.; Werwatz, A.; Kraehnert, R. *Nat Commun* 2019.



### Elemental composition of catalyst (mol%)

## Process parameters + Preparation      Catalytic performance

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	No. of publications	Cation 1 value	Cation 2 mol%	Cation 3 mol%	Cation 4 mol%	Anion 1 value	Anion 2 value	Preparation	Temperature, K	p(CH <sub>4</sub> )/bar	p(O <sub>2</sub> )/bar	p(CH <sub>4</sub> )/p(O <sub>2</sub> )	Reaction time, s	X(O <sub>2</sub> ), %	X(CH <sub>4</sub> ), %	S(C <sub>1</sub> ), %	S(C <sub>2</sub> ), %	S(C <sub>3</sub> ), %	Y(C <sub>1</sub> ), %	Y(C <sub>2</sub> ), %	Y(C <sub>3</sub> ), %											
2		mol%	value	mol%	value	mol%	value																									
212		Li	22.2	Ca	77.8														n.a.	984	0.47	0.11	4.4	1.0	8.70		18.0	33.0	38.0	27.0	65.0	11.7
213		Na	7.3	Ca	92.7														n.a.	982	0.47	0.11	4.4	1.0	8.70		15.0	25.0	35.0	26.0	55.0	11.2
214		Li	17.2	Mg	82.8														n.a.	1043	0.38	0.13	3.0	1.0	5.50		59.0	48.0	36.0	17.0	53.0	31.3
215		Sn	1.2	Li	17.6	Mg	81.20												n.a.	973	0.47	0.11	4.4	1.0	7.50		23.0	31.0	35.0	27.0	62.0	14.3
216	23	Li	22.2	Ca	77.8														n.a.	984	0.47	0.11	4.4	1.0	8.70		18.0	33.0	38.0	27.0	65.0	11.7
217		Na	7.3	Ca	92.7														n.a.	982	0.47	0.11	4.4	1.0	8.70		15.0	25.0	35.0	26.0	55.0	11.2
218		Na	8.5	Pb	8.5	Si	8.50	Zn	74.50										n.a.	1023	0.90	0.13	9.0	1.0	2.40	71.0	8.0	34.0	34.0	68.0	5.4	
219		Na	8.0	K	4.0	Ce	20.00	Zn	50.00	S	16.0	P	2.0	C				Therm.decomp.	970	0.50	0.11	4.8	1.0	1.20		19.2		43.2	17.6	66.8	11.7	
220		Zr	100.0																n.a.	973	0.47	0.11	4.2	1.0	1.20		12.0	0.0	2.5	2.5	0.3	
221		Na	8.5	Pb	8.5	Si	8.50	Zn	74.50										n.a.	1023	0.90	0.13	9.0	1.0	2.40	71.0	8.0	34.0	34.0	68.0	5.4	
222		Ca	100.0																n.a.	1023	0.91	0.09	10.0	1.0	2.00		5.0	30.0	54.0	84.0	4.2	
223		Li	100.0																n.a.	1023	0.91	0.09	10.0	1.0	2.00		5.0	44.0	44.0	50.0	4.5	
224		Li	33.3	Ca	33.3	Pr	33.40												n.a.	1023	0.91	0.09	10.0	1.0	2.00		14.0	71.0	13.0	84.0	11.8	
225		Li	33.3	Ca	33.3	Ce	33.40												n.a.	1023	0.91	0.09	10.0	1.0	2.00		14.0	70.0	14.0	84.0	11.8	
226		Ca	100.0																n.a.	1023	0.67	0.07	10.0	1.0	2.00		14.0	71.0	11.0	82.0	11.5	
227		Mg	100.0																n.a.	1023	0.67	0.07	10.0	1.0	2.00		5.0	38.0	46.0	57.0	4.4	
228		Mn	100.0																n.a.	1023	0.67	0.07	10.0	1.0	2.00		14.0	71.0	13.0	84.0	11.8	
229		Pb	100.0																n.a.	1023	0.67	0.07	10.0	1.0	2.00		14.0	48.0	18.0	66.0	6.5	

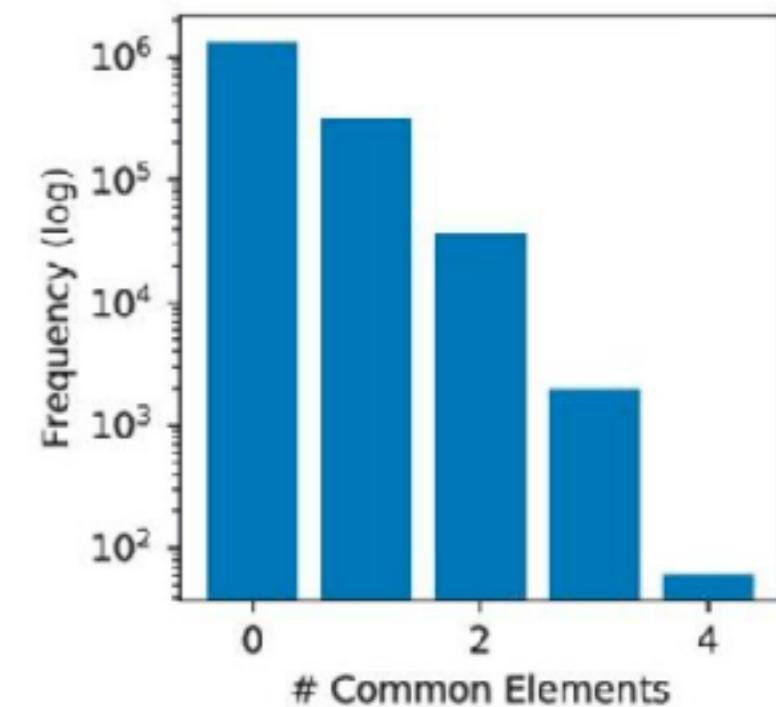
# 文献から収集したデータの問題

- ✓ **Underspecification:** 実験条件が強く影響するが、複数の実験条件で報告例のある触媒は極めて少ない。同一条件での複製実験もなし。  
→ 1866例中、2条件以上 158例、3条件以上 60例、4条件以上 26例
- ✓ **Sparsity:** 組成で使われる元素のオーバラップが少なく非常にスパース  
→ 例えば '[Na:33.2 Ti:0.5 Mn:66.3](#)' と '[Zn:77.8 Ce:22.2](#)' をどう比較すべき？

74 elements

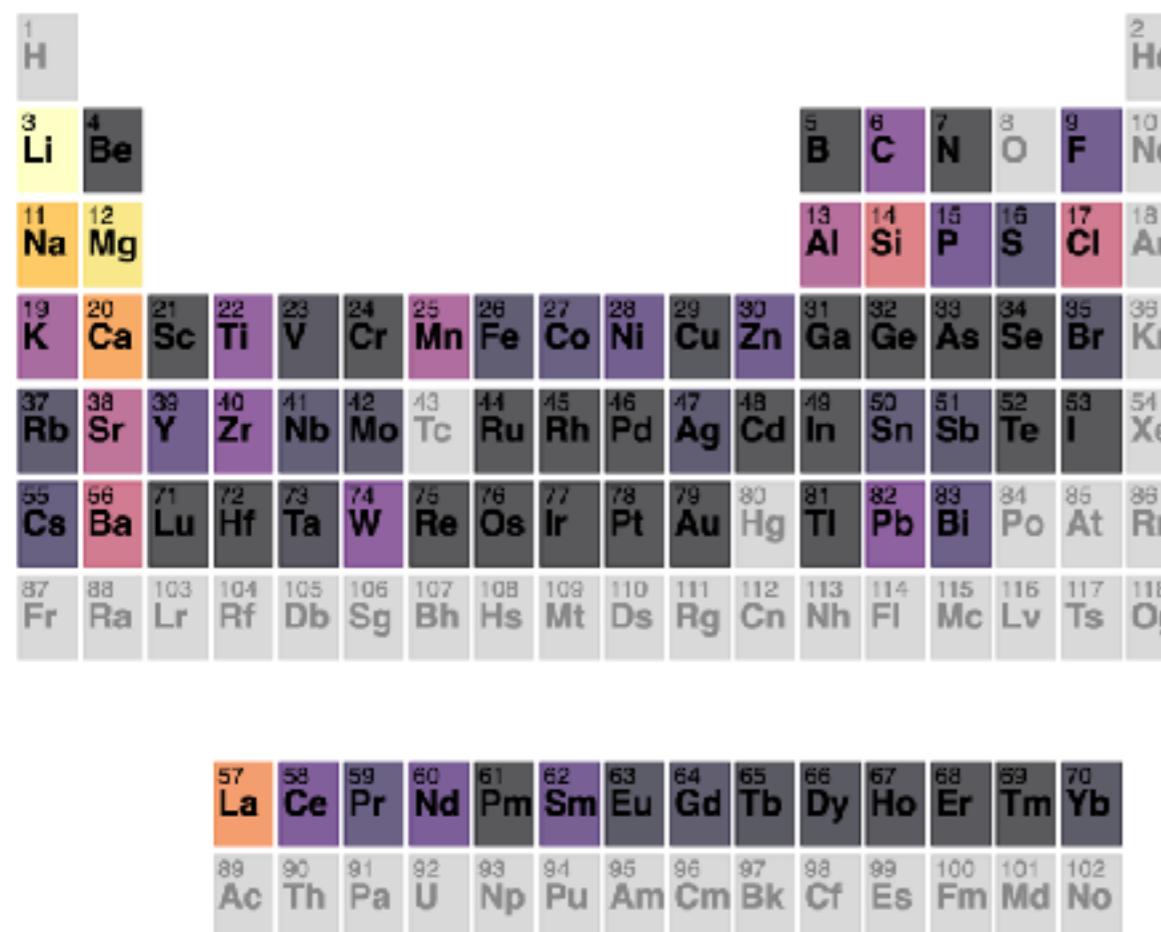
Li	Be	B	C	N	F	Na	Mg	Al	Si	...	Ta	W	Re	Os	Ir	Pt	Au	Tl	Pb	Bi
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	90.000003	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	95.300003	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	95.500000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	89.599998	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	97.199997	0.0	...	0.0	0.0	0.0	0.0	0.0	2.8	0.0	0.0	0.0	0.000000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
0.0	0.0	0.0	0.0	0.0	0.0	23.799999	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.799999	
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	

All pairwise comparisons

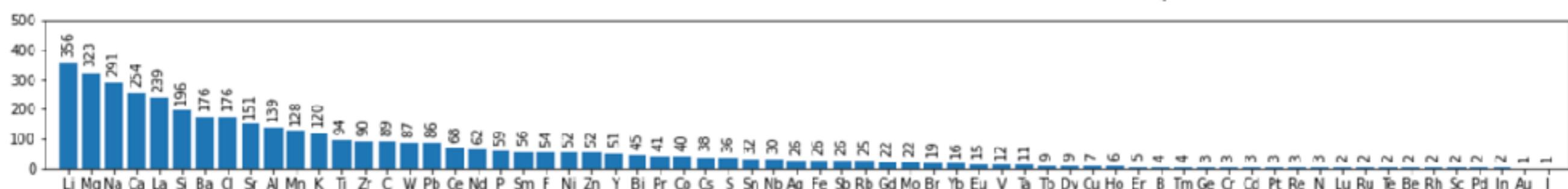
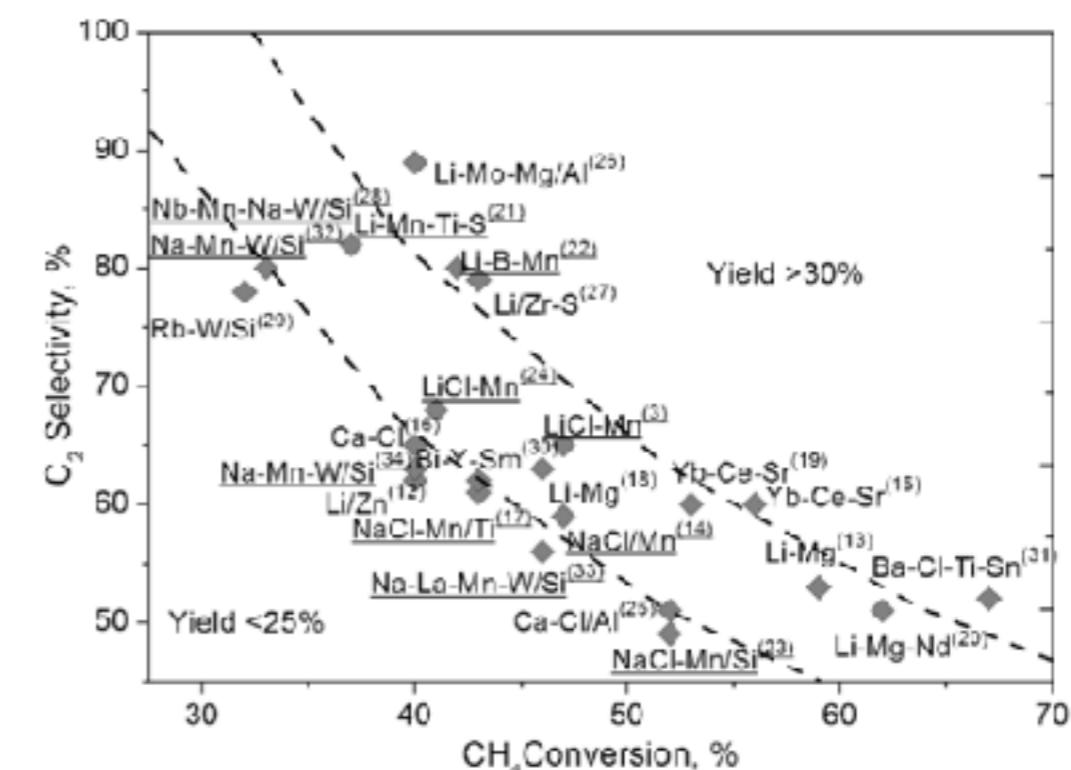


# 文献から収集したデータの問題

- ✓ **Strong Bias:** 出版される科学成果の強い成功バイアス、流行や実験のしやすさによる選択バイアス、など認知バイアス・社会的バイアスの影響



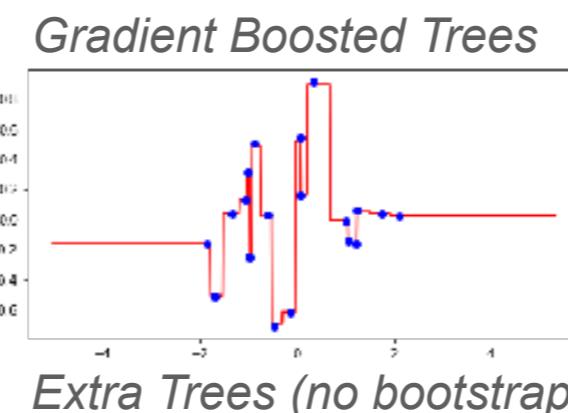
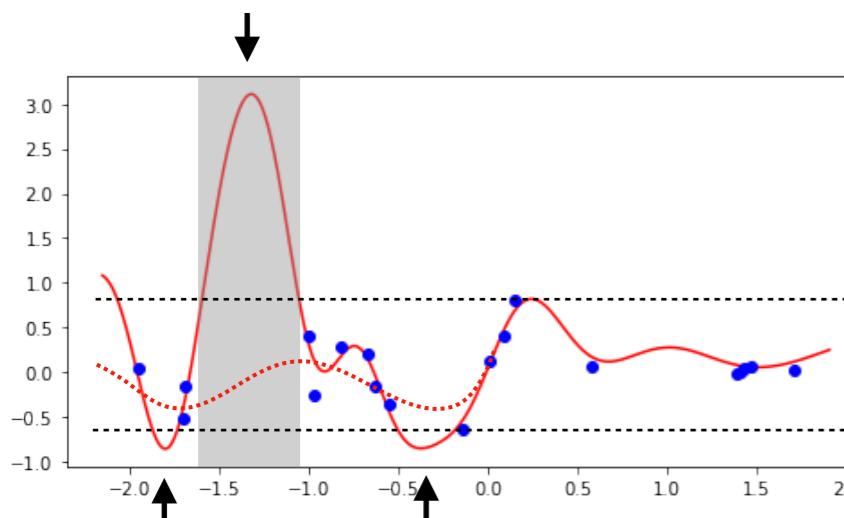
LaO<sub>3</sub>, Li/MgO, Mn/Na<sub>2</sub>WO<sub>4</sub>/SiO<sub>2</sub>など  
特定の触媒が非常によく研究される



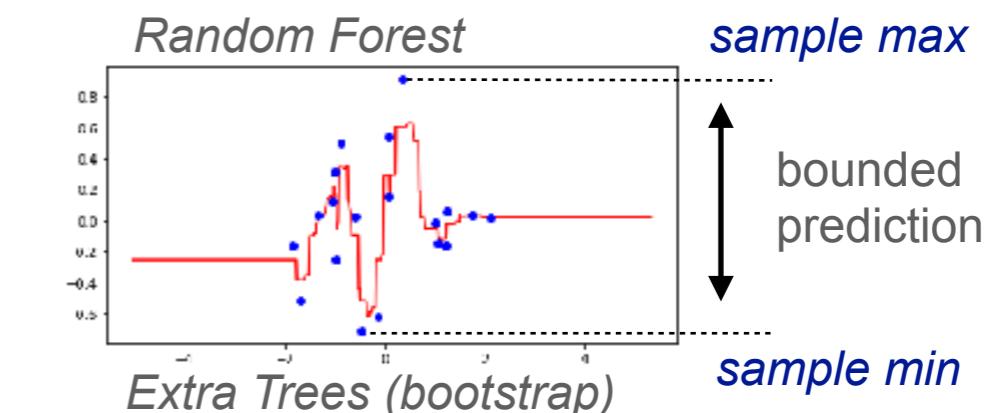
# 機械学習予測の確度・信頼度も見て適用範囲を慎重に理解

## ✓ 決定木アンサンブルの予測分散 (信頼区間)

訓練データのmax-minの間に予測値があるので意図しない外挿リスクが少ない



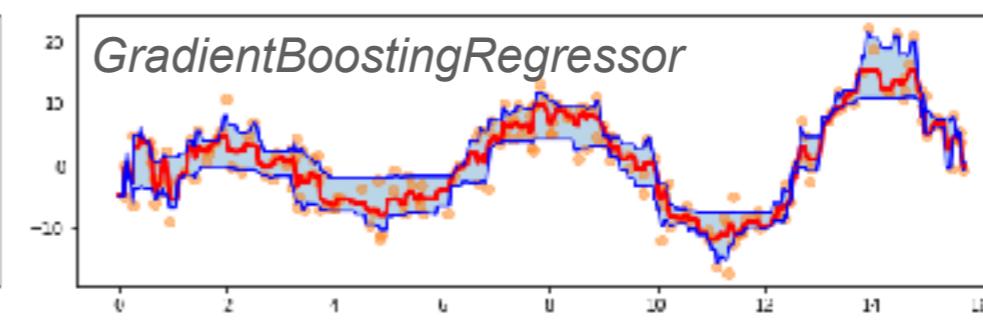
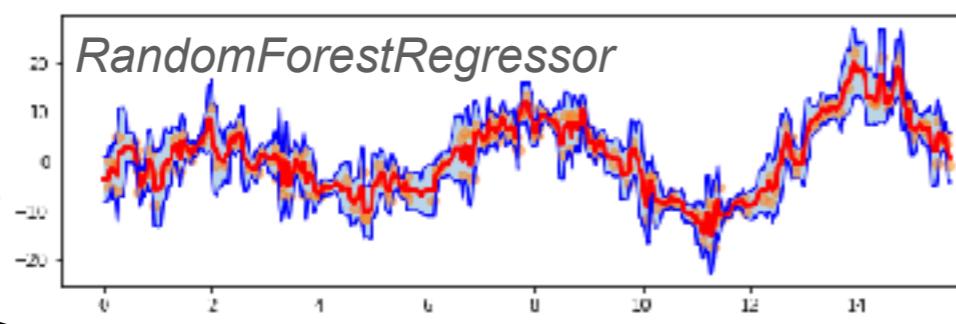
*Extra Trees (no bootstrap)*



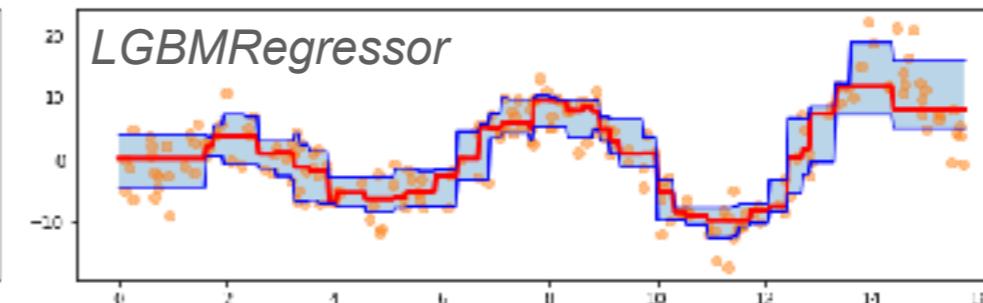
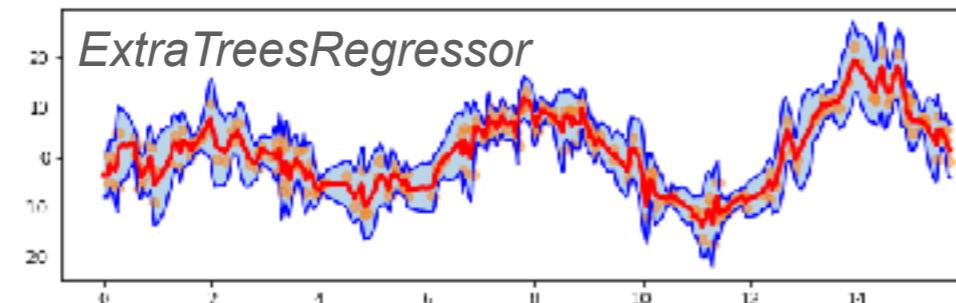
*Extra Trees (bootstrap)*

sample max  
↓  
bounded prediction  
↑ sample min

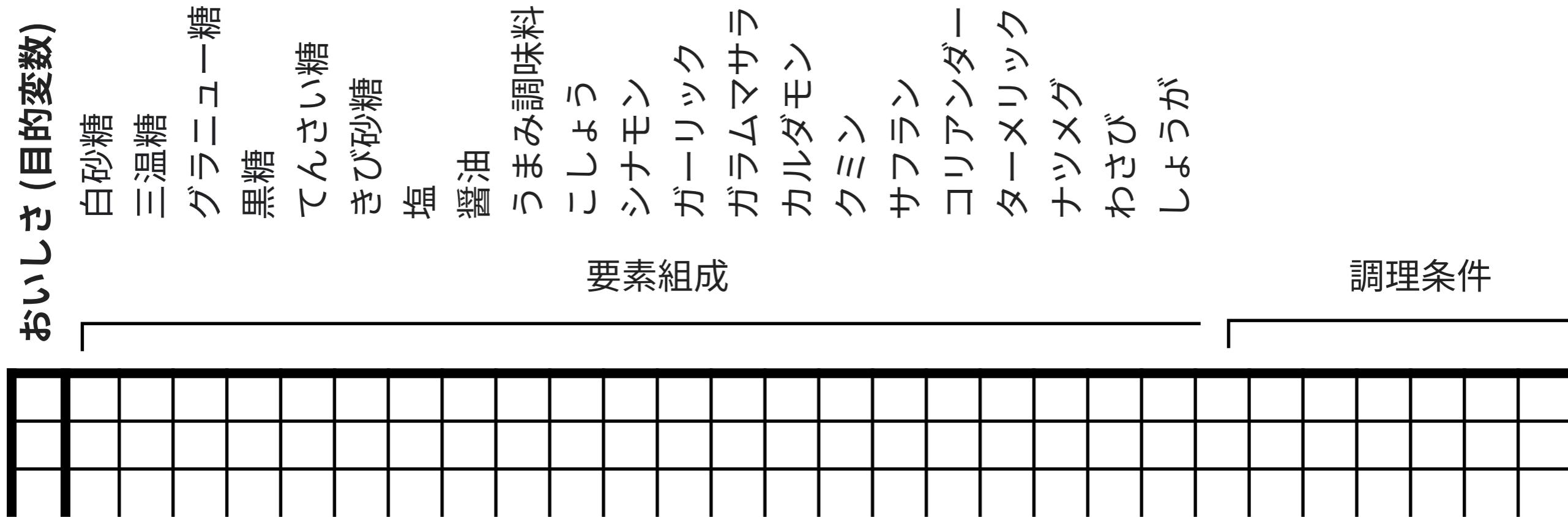
Naturally by the law of total variance



By quantile regression to .16, .5, .84 quantiles



# 特徴量の設計：触媒の効果的な特徴表現を考える



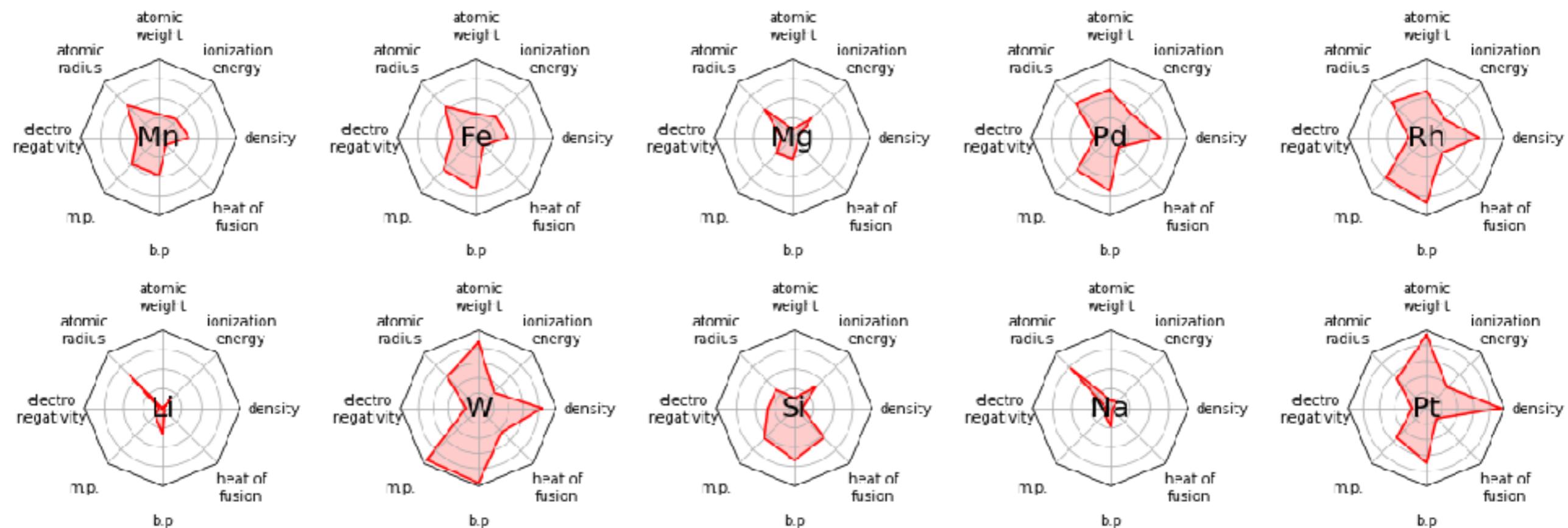
- このままでは、白砂糖～きび砂糖まではショ糖で「甘味」、醤油には「塩分」が含まれる、など、要素の「個性」は全く考慮されない。
- 訓練データに含まれない要素が入ると予測に反映できない。
- 要素ごとの頻度がべき乗則的であり非常に大きな偏りがある。
- 要素数が多く報告例に要素のオーバラップが少ない。

# 元素の個性を元素記述子ベクトルで表す入力表現

元素を「シンボル」として扱うのではなく  
「多次元の元素記述子ベクトル」で扱う

元素記述子の抽象度を変えれば、関心のある特性のみに着目して元素の表現・比較が可能に  
**(訓練データに含まれない元素も扱える！)**

Element	Descriptors			
	1	2	...	p
A	A <sub>1</sub>	A <sub>2</sub>	...	A <sub>p</sub>
B	B <sub>1</sub>	B <sub>2</sub>	...	B <sub>p</sub>
C	C <sub>1</sub>	C <sub>2</sub>	...	C <sub>p</sub>
D	D <sub>1</sub>	D <sub>2</sub>	...	D <sub>p</sub>
E	E <sub>1</sub>	E <sub>2</sub>	...	E <sub>p</sub>



# 元素の個性を元素記述子ベクトルで表す入力表現

## SWED (Sorted Weighted Elemental Descriptors) 表現：

組成比 × 元素記述子ベクトルを組成比の降順に並べたもの  
→ シンプルだが定量的な改善が得られた特徴ベクトル表現



Literature data

Catalyst	Composition [mol%]				
	A	B	C	D	E
Cat-ABC1	90	4	6	0	0
Cat-BDE1	0	11	0	80	9
Cat-AE1	75	0	0	0	25
Cat-AE2	80	0	0	0	20
Cat-ABCDE1	2	3	15	10	70

Element	Descriptors			
	1	2	...	p
A	A <sub>1</sub>	A <sub>2</sub>	...	A <sub>p</sub>
B	B <sub>1</sub>	B <sub>2</sub>	...	B <sub>p</sub>
C	C <sub>1</sub>	C <sub>2</sub>	...	C <sub>p</sub>
D	D <sub>1</sub>	D <sub>2</sub>	...	D <sub>p</sub>
E	E <sub>1</sub>	E <sub>2</sub>	...	E <sub>p</sub>

Descriptors	8 descriptors (p=8)	3 descriptors (p=3)
AW	○	
Atomic radius	○	
Electronegativity	○	○
m.p.	○	
b.p.	○	
$\Delta H_{us}$	○	○
Density	○	○
Ionization energy	○	

**Sorted Weighted Elemental Descriptor (SWED) representation**  
K is the max number of elements in a catalyst. K = 5 in this example, K = 8 used in this paper

Catalyst	1 <sup>st</sup> feature				2 <sup>nd</sup> feature				...	K <sup>th</sup> feature			
	1	2	...	p	1	2	...	p	...	1	2	...	p
Cat-ABC1	90 × A <sub>1</sub>	90 × A <sub>2</sub>	...	90 × A <sub>p</sub>	6 × C <sub>1</sub>	6 × C <sub>2</sub>	...	6 × C <sub>p</sub>	...	0 × E <sub>1</sub>	0 × E <sub>2</sub>	...	0 × E <sub>p</sub>
Cat-BDE1	80 × D <sub>1</sub>	80 × D <sub>2</sub>	...	80 × D <sub>p</sub>	11 × B <sub>1</sub>	11 × B <sub>2</sub>	...	11 × B <sub>p</sub>	...	0 × C <sub>1</sub>	0 × C <sub>2</sub>	...	0 × C <sub>p</sub>
Cat-AE1	75 × A <sub>1</sub>	75 × A <sub>2</sub>	...	75 × A <sub>p</sub>	25 × E <sub>1</sub>	25 × E <sub>2</sub>	...	25 × E <sub>p</sub>	...	0 × D <sub>1</sub>	0 × D <sub>2</sub>	...	0 × D <sub>p</sub>
Cat-AE2	80 × A <sub>1</sub>	80 × A <sub>2</sub>	...	80 × A <sub>p</sub>	20 × E <sub>1</sub>	20 × E <sub>2</sub>	...	20 × E <sub>p</sub>	...	0 × D <sub>1</sub>	0 × D <sub>2</sub>	...	0 × D <sub>p</sub>
Cat-ABCDE1	90 × E <sub>1</sub>	90 × E <sub>2</sub>	...	90 × E <sub>p</sub>	15 × C <sub>1</sub>	15 × C <sub>2</sub>	...	15 × C <sub>p</sub>	...	2 × A <sub>1</sub>	2 × A <sub>2</sub>	...	2 × A <sub>p</sub>

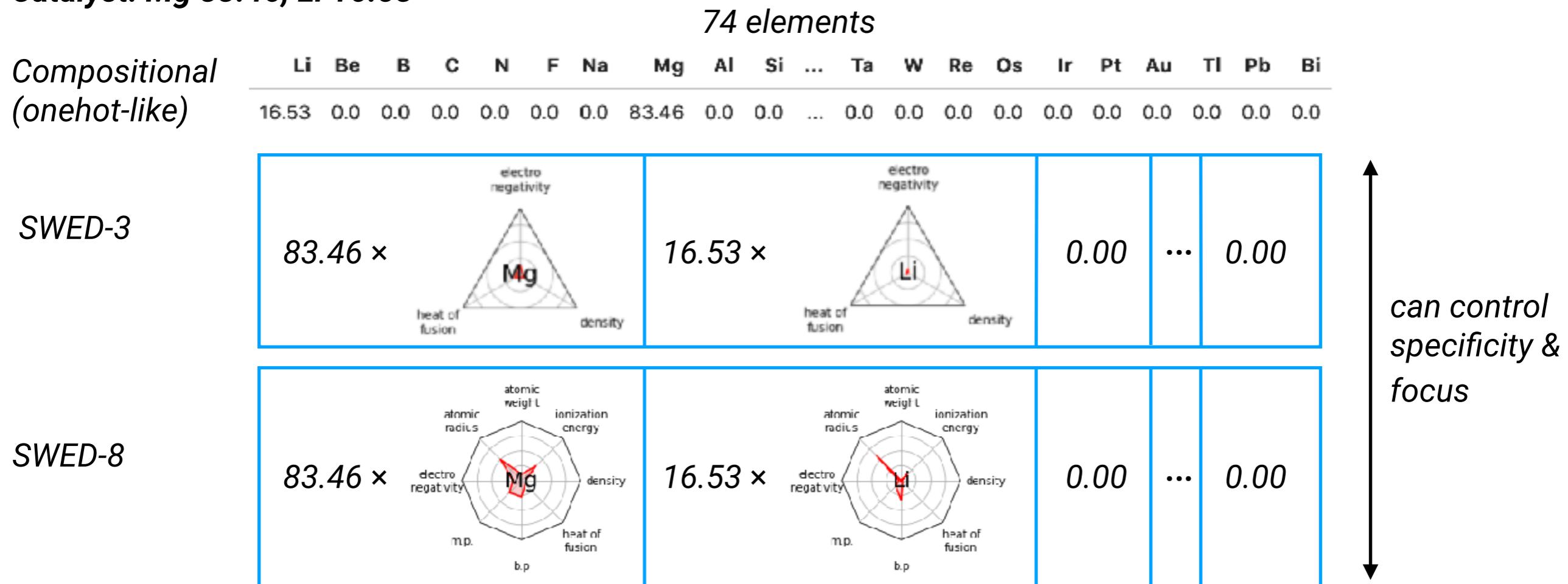
+

Experimental condition				
Imp.	SG	Pre	Temp.	$P_{\text{CH}_4}/P_{\text{O}_2}$
1	0	0	1023	3
0	1	0	1023	2.5
0	0	0	923	3
0	0	1	923	5
1	0	0	973	4

# SWED表現と用いる元素記述子による抽象度の制御

- ✓ 元素記述子としてどのようなものを使うかによって粗視化が制御可能
- ✓ 元素は選んだ元素記述子の数値で表現され、この表現のもとで内挿されるため、訓練データにはない元素も自然に取り扱うことができる。

**Catalyst: Mg 83.46, Li 16.53**



SWED-3 features: electronegativity, density, enthalpy of fusion

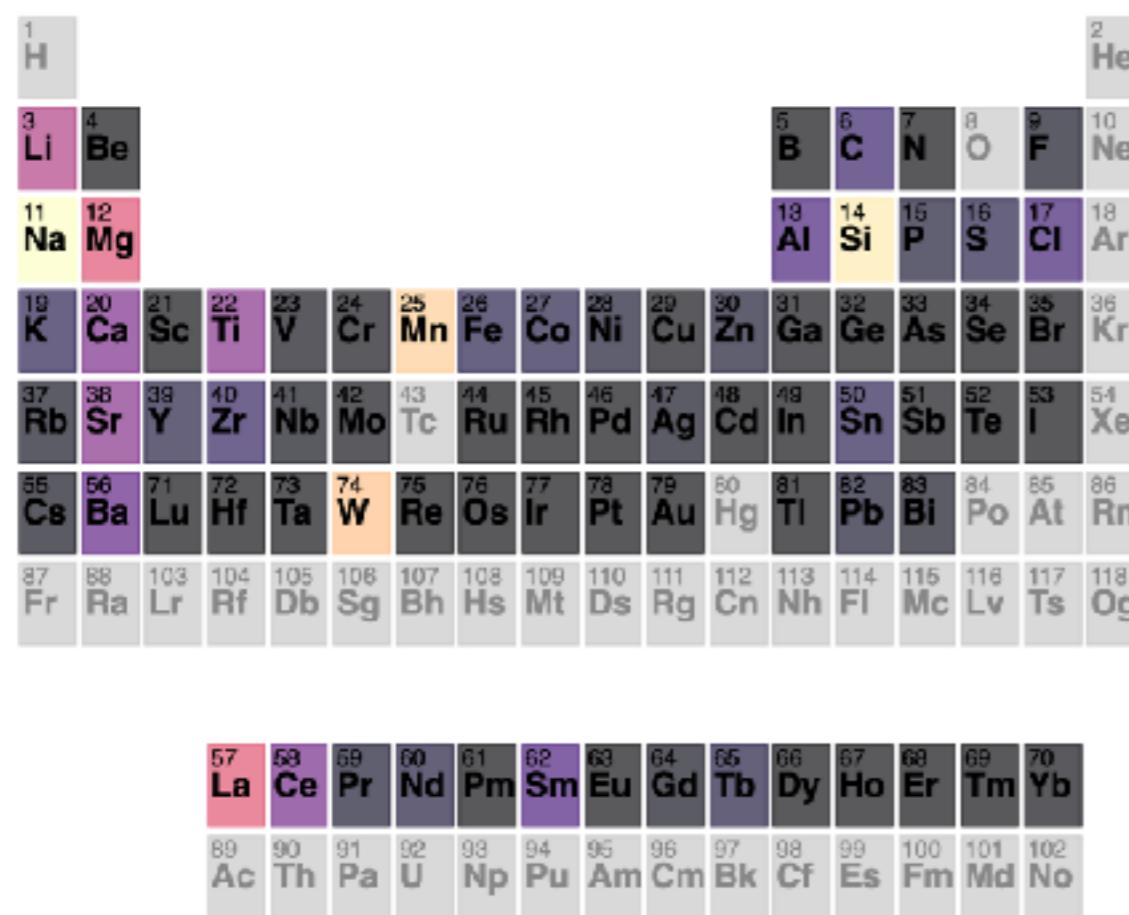
SWED-8 features: SWED-3 features + atomic weight, atomic radius, m.p., b.p., ionization energy

# 新しいトレンドを反映するためデータセット自体も拡充

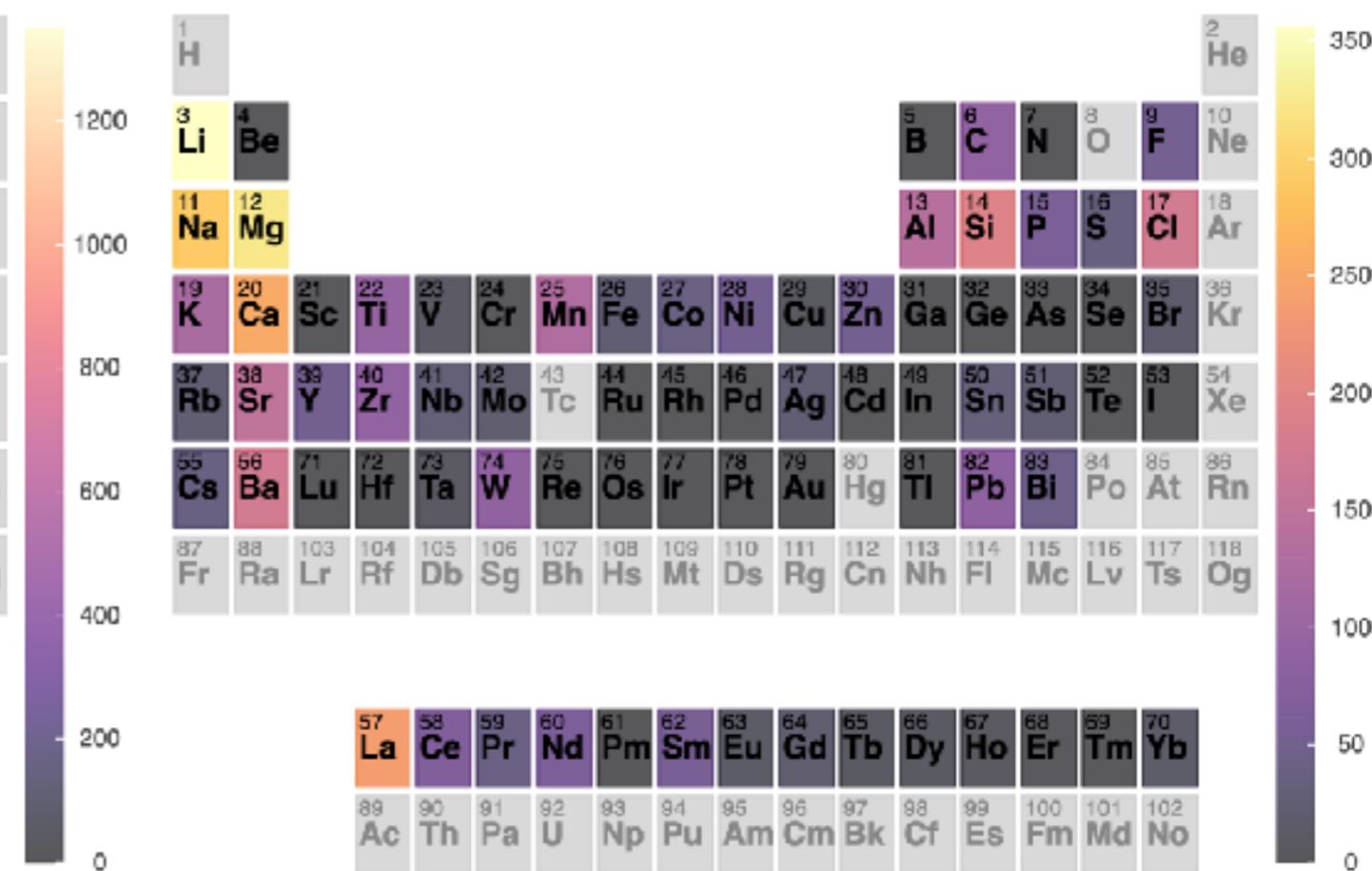
## 4559 catalyst records from 542 reports

Mine, S.; Takao, M.; Yamaguchi, T.; Toyao, T.\*; Maeno, Z.; Hakim Siddiki, S. M. A.; Takakusagi, S.; Shimizu, K.\*; Takigawa, I.\* *ChemCatChem* 2021. <https://doi.org/10.1002/cctc.202100495>.

The update dataset:  
**4559 catalyst records**  
**from 542 reports (2010 - 2019)**



The original dataset:  
1866 catalyst records  
from 421 reports (1982 - 2009)



# C<sub>2</sub>収率の機械学習予測

1. **Conventional:** composition + condition
2. **Proposed(Exploitative):** composition + SWED + condition
3. **Proposed(Explorative):** SWED + condition w/ SWED → composition estimator

**Table 2.** Comparison of prediction accuracy (RMSE) and coefficient of determination ( $R^2$ ) for the C<sub>2</sub> yields (%) of the OCM reaction. Two datasets and three ML methods were tested using 10-fold cross-validation. The numbers shown in parentheses are the corresponding  $\sigma$ s.

	Pre-2010 dataset			Entire OCM dataset		
ML model	RFR	ETR	XGB	RFR	ETR	XGB
<b>Conventional Method</b>						
Training Error [%]	1.66 (0.02)	0.17 (0.03)	1.07 (0.37)	1.50 (0.02)	0.75 (0.03)	2.21 (0.38)
Test Error [%]	4.50 (0.38)	4.65 (0.50)	4.34 (0.34)	3.66 (0.23)	3.65 (0.20)	3.71 (0.23)
Test R <sup>2</sup>	0.536	0.504	0.567	0.713	0.716	0.706
<b>Proposed Method (Exploitative)</b>						
Training Error [%]	1.63 (0.02)	0.17 (0.02)	0.55 (0.31)	1.50 (0.02)	0.76 (0.03)	1.73 (0.26)
Test Error [%]	4.39 (0.43)	4.30 (0.52)	4.25 (0.41)	3.66 (0.27)	3.52 (0.25)	3.58 (0.28)
Test R <sup>2</sup>	0.557	0.575	0.583	0.713	0.736	0.722
<b>Proposed Method (Explorative) with all the 8 descriptors</b>						
Training Error [%]	1.68 (0.02)	0.17 (0.02)	0.29 (0.10)	1.52 (0.02)	0.76 (0.03)	1.32 (0.31)
Test Error [%]	4.50 (0.48)	4.44 (0.50)	4.43 (0.52)	3.70 (0.29)	3.57 (0.27)	3.56 (0.28)
Test R <sup>2</sup>	0.536	0.547	0.547	0.708	0.727	0.728
<b>Proposed Method (Explorative) with 3 descriptors<sup>[a]</sup></b>						
Training Error [%]	1.66 (0.02)	0.17 (0.02)	0.34 (0.14)	1.52 (0.02)	0.76 (0.03)	1.27 (0.18)
Test Error [%]	4.45 (0.34)	4.45 (0.35)	4.41 (0.35)	3.69 (0.30)	3.63 (0.26)	3.56 (0.27)
Test R <sup>2</sup>	0.547	0.540	0.556	0.709	0.717	0.728

[a] The electronegativity, density, and  $\Delta H_{fus}$  were used as descriptors.

RFR (Random Forest); ETR (ExtraTrees); XGB (XGBoost)

SWED-3 features: electronegativity, density, enthalpy of fusion

SWED-8 features: SWED-3 features + atomic weight, atomic radius, m.p., b.p., ionization energy

# SWED-3を用いた触媒候補 (期待改善値で上位20個)

Table 3. 20 most promising candidate catalyst systems for further testing in the OCM, as suggested using SMBO with ETR coupled with the proposed method (explorative) and the entire dataset. In addition to Els, the predicted values  $\mu$ , standard deviations  $\sigma$ , and 95% confidence intervals (as  $\mu \pm 1.96\sigma$ ) are also shown. Oxygen is not shown in the elemental compositions. Only three descriptors, electronegativity, density, and  $\Delta H_{fu}$ , were used.

Elemental composition	Promoter	Preparation method	T [K]	$P(\text{CH}_4)/P(\text{O}_2)$	$P_{\text{total}}$ [bar]	Contact time [s]	El	Predicted C <sub>x</sub> yield [%]			
								mean	sd	95%CI lower	95%CI upper
Mn:72.3 Li:27.7	B	Solid-phase technique	1023	1.67	1.01	1.20	2.29	25.69	13.52	-0.80	52.19
Sr:50.0 Ce:45.0 Yb:5.0	-	Solid-phase technique	1023	1.99	1.01	0.79	2.29	25.88	13.35	-0.28	52.04
Si:60.9 Na:19.3 Cl:17.2 Mn:1.6 W:1.0	-	Hydrothermal treatment	1023	1.60	1.01	0.02	1.67	25.55	11.71	2.60	48.50
Mn:80.0 Li:20.0	Cl	Solid-phase technique	1023	1.96	1.00	0.60	1.37	23.63	12.12	-0.13	47.40
Mg:82.8 Li:17.2	-	Solid-phase technique	1043	3.00	1.01	5.50	1.31	22.97	12.38	-1.29	47.23
C:42.7 Sc:23.6 Ge:17.8 K:10.1 I:5.8	-	Physical mixing	1023	1.60	1.01	0.00	1.29	24.00	11.56	1.34	46.66
Si:45.9 Mg:22.0 Ru:20.5 Sc:6.0 Ge:5.7	-	Physical mixing	1023	1.62	1.39	0.03	1.22	23.48	11.69	0.57	46.39
Ge:30.9 Sc:30.7 As:25.3 Be:6.7 I:6.3	-	Physical mixing	1023	1.41	2.75	0.13	1.05	22.75	11.56	0.11	45.40
Si:35.5 Br:32.4 Mg:14.4 Al:9.0 Ho:6.5 Y:2.1	S	Physical mixing	1023	1.28	1.01	0.00	1.01	25.54	9.49	5.94	44.13
Mg:36.0 Ge:33.6 Mo:30.3	-	Precipitation	1073	2.50	1.01	3.60	0.95	23.97	10.34	3.71	44.23
La:46.4 Ge:27.9 Cu:25.7	-	Ceramic method	1023	1.94	0.68	2.40	0.95	23.16	10.86	1.88	44.44
Sc:36.9 Ca:32.9 Mo:30.2	-	Ceramic method	1040	0.90	0.70	1.79	0.92	23.84	10.31	3.63	44.05
Nd:83.6 Ge:16.4	-	Hydrothermal treatment	1100	3.95	2.24	0.35	0.92	22.20	11.38	-0.10	44.50
Si:34.4 Ca:29.1 Ge:23.2 Nb:9.7 As:3.6	-	Physical mixing	1002	1.39	1.90	0.02	0.91	20.33	12.55	-4.26	44.92
C:45.8 Sc:20.8 Ge:20.3 Mo:7.7 Nb:5.3	-	Physical mixing	1015	0.89	2.96	0.04	0.91	20.79	12.22	-3.16	44.75
Sc:49.4 Au:34.1 Ge:16.5	-	Ceramic method	1023	2.00	1.01	2.00	0.90	22.62	11.00	1.06	44.19
C:38.8 Sc:31.5 Ge:16.0 As:7.9 Rh:5.8	-	Physical mixing	1017	0.63	3.02	0.07	0.89	21.32	11.83	-1.88	44.51
Mo:38.9 V:37.9 Ge:23.2	-	Ceramic method	1048	2.00	0.85	2.03	0.89	22.39	11.14	0.57	44.22
Sr:45.7 Ge:33.7 As:20.6	-	Ceramic method	1023	1.99	0.69	1.20	0.89	22.39	11.14	0.56	44.22
Sc:38.7 Mo:37.5 Ca:23.9	-	Ceramic method	1023	1.96	1.01	4.10	0.88	21.29	11.80	-1.85	44.42

訓練データにない元素  
(As, Hf, Se, Os, Pm)も提  
案候補に見られた

毒性のため実用上は問  
題があるものの、そ  
のような候補もきちんと  
探索されている！

# 学習には「知識の利用」と「探索」のトレードオフが伴う

新しいことを「学ぶ」際の最も基本的なトレードオフ

## 1. 今まで学んだことの「利用」

今までのデータの機械学習に基づく予測の活用

※ 全体に占める「今までに学んだこと」のカバー率が  
低い場合は木を見て森を見ずになってしまう

## 2. 今までに学んでないことの「探索」

= 新しい経験、新しい知識の吸収、知識の拡充

今までのデータの確度が低い領域、データがない領域からの  
更なるデータの取得！

# 文献データには様々な問題があり「探索」がより重要

- ✓ 実験は人間が計画するため、認知バイアスや社会的バイアスが反映されてしまう(従来知見、流行、伝統、実験しやすさ…)
- ✓ 訓練データの事例の分布に大きな偏りがあるが、これは自然の摂理ではなく私たちの視野の狭さ(思い込み)を反映したもの
- ✓ マタイ効果 (Matthew effect)：成功例に過剰に引きずられがち
- ✓ 成功例のみが報告されるため失敗事例の情報が致命的に欠損

LETTER

12 SEPTEMBER 2019 | VOL 573 | NATURE | 251

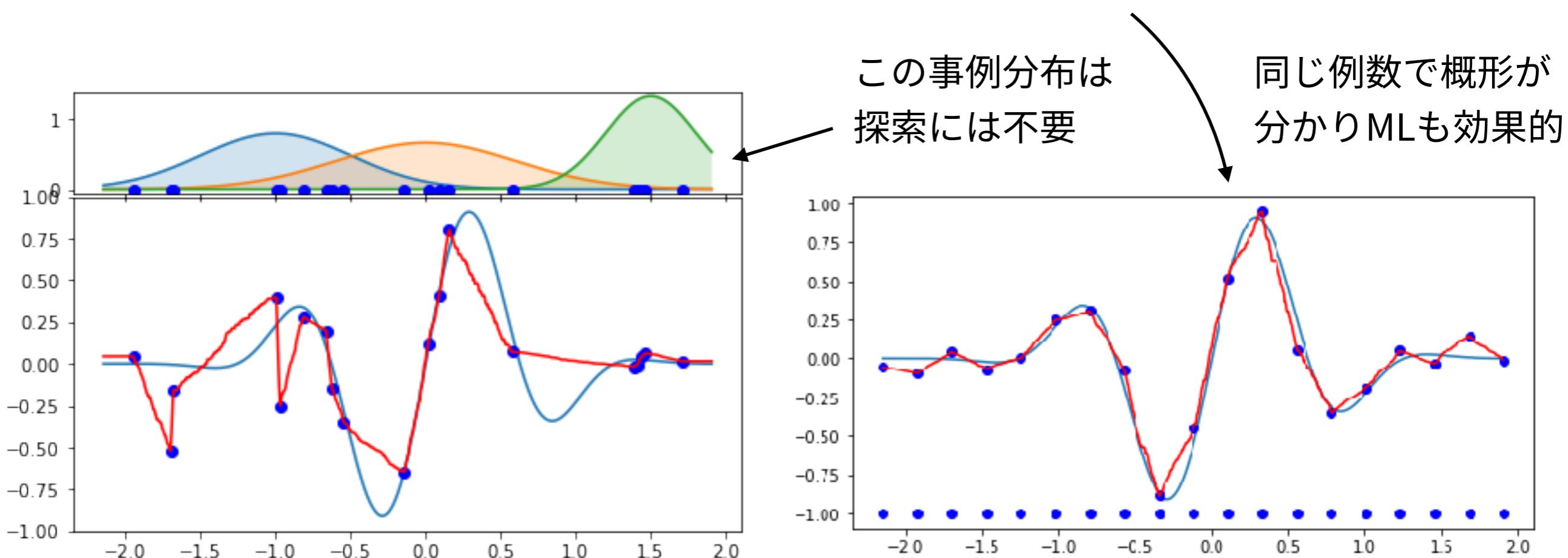
<https://doi.org/10.1038/s41586-019-1540-5>

## Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis

Xiwen Jia<sup>1</sup>, Allyson Lynch<sup>1</sup>, Yuheng Huang<sup>1</sup>, Matthew Danielson<sup>1</sup>, Immaculate Lang'at<sup>1</sup>, Alexander Milder<sup>1</sup>, Aaron E. Ruby<sup>1</sup>, Hao Wang<sup>1</sup>, Sorelle A. Friedler<sup>2\*</sup>, Alexander J. Norquist<sup>1,\*</sup> & Joshua Schrier<sup>1,3\*</sup>

# 実験計画と機械学習

- 探索点を計画できる場合は、**生起想定範囲にできるだけ「まんべんなく」とる**ほうが良い (e.g. ランダム実験、完全実施要因計画、ラテン超方格計画、D最適計画、…)
- 探索や実験計画においてはMLは現実の代理モデルにすぎない



# 実験計画におけるフィッシュナーの三原則

# 実験計画におけるフィッシャーの三原則

## 1. 反復 (replication)

→ 同条件で複数回の実験を行う。再現性の担保に加え、この情報がないと系統誤差と偶然誤差を判別できない。

# 実験計画におけるフィッシャーの三原則

## 1. 反復 (replication)

→ 同条件で複数回の実験を行う。再現性の担保に加え、この情報がないと系統誤差と偶然誤差を判別できない。

## 2. 無作為化 (randomization)

→ 考えたい要因以外に目的変数に影響を与える可能性がある要因がある場合、可能な限りランダムに割り付けする。

(c.f. 結果がよかった条件まわりで多めに試したいのは理解できるが探索が目的ならむしろランダム実験条件のほうが良い)

# 実験計画におけるフィッシャーの三原則

## 1. 反復 (replication)

→ 同条件で複数回の実験を行う。再現性の担保に加え、この情報がないと系統誤差と偶然誤差を判別できない。

## 2. 無作為化 (randomization)

→ 考えたい要因以外に目的変数に影響を与える可能性がある要因がある場合、可能な限りランダムに割り付けする。

(c.f. 結果がよかった条件まわりで多めに試したいのは理解できるが探索が目的ならむしろランダム実験条件のほうが良い)

## 3. 局所管理 (local control)

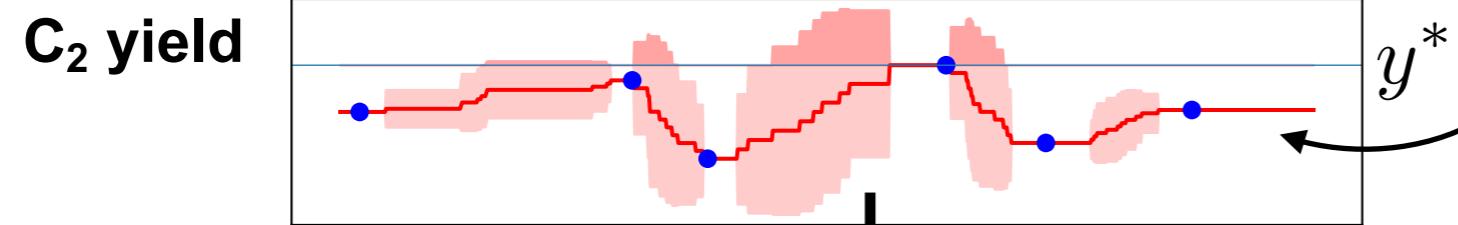
→ 考えたい要因以外のバックグラウンド因子はできるだけ均一になるように実験を管理する。

(c.f. 実験条件の最適化は諦めて固定し組成だけふる実験)

# 実験計画と機械学習

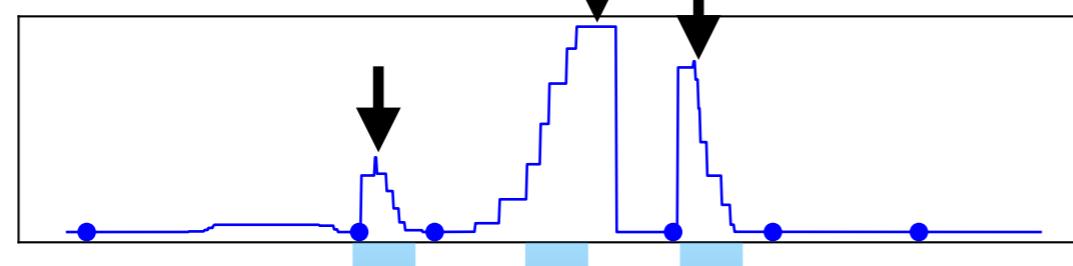
ML surrogate w/ 95%CI  $\mu(x) \pm 1.96\sigma(x)$

- $n$  given data points



良い特徴表現と予測精度の高い**代理モデル**

Expected Improvement



予測分布を考慮した**探索指標**の局所ピーク ↓ 同定

似た候補点はグループ化し代表点を表示

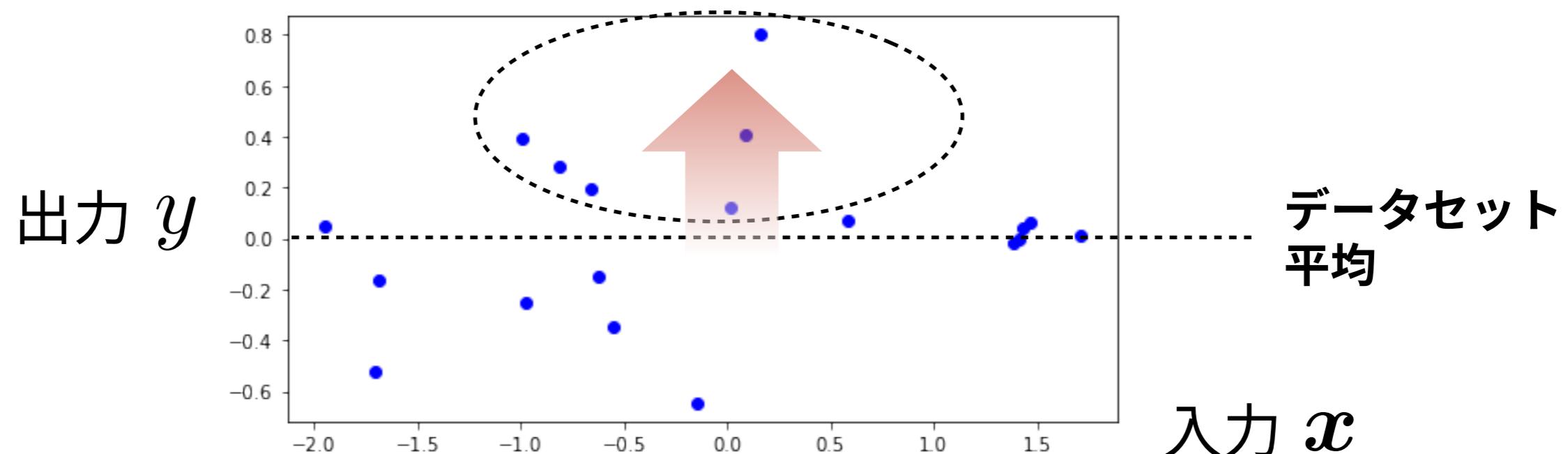
次に実験する価値の高い候補点を diversify し提示

MLに入力するデータにない傾向は原理上予測できないため  
アルゴリズムの詳細よりも**データの収集計画 (実験計画)、適用範囲の理解、品質保証**が成功の鍵であることをいつも心に

# 機械学習モデルがなぜその予測をしたかの要因分析



収率  $y$  が高い触媒と低い触媒の違いを規定しうる因子は何だろう？(ただし入力  $x$  が含む情報の範囲で)

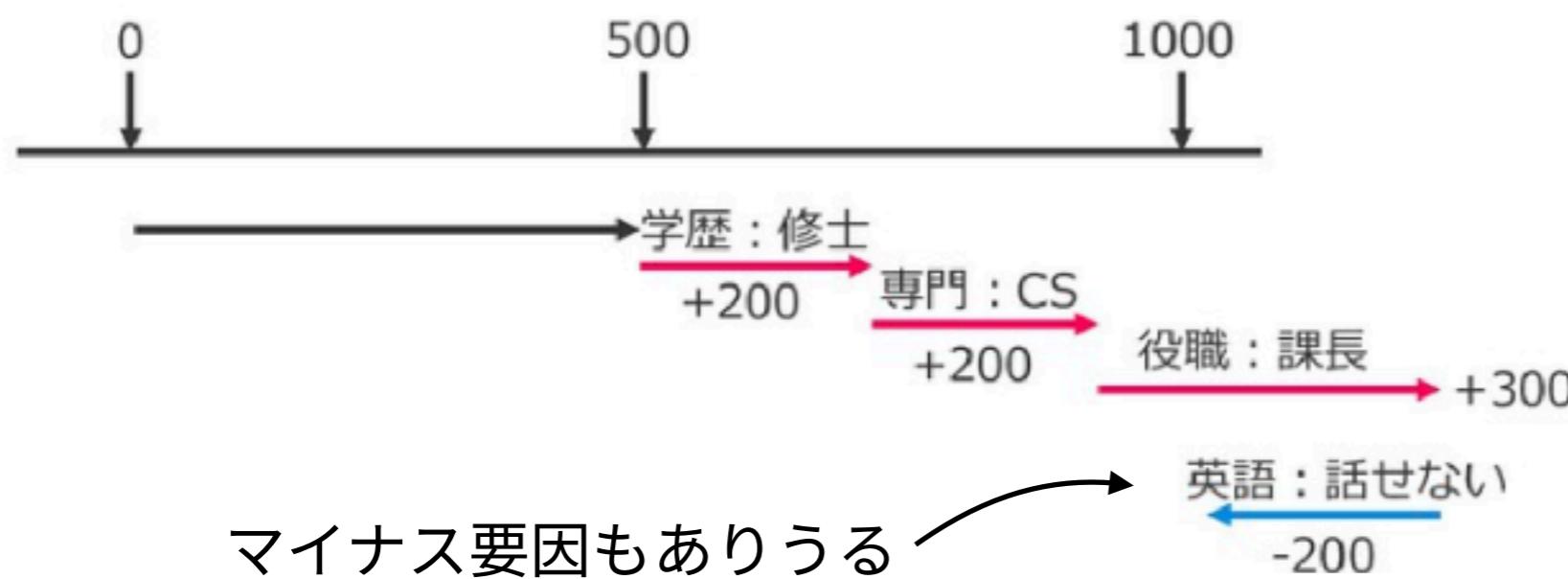


# 機械学習モデルがなぜその予測をしたかの要因分析

## SHAP (SHapley Additive exPlanations)

与えられた予測値のデータセット平均からの変化量を  
**「特徴量ごとの寄与度(SHAP値)の和」**へ分解するモデル説明法

予測の平均値は年収500万なのにこの個人は年収1000万と予測された  
 → 500万の差分はどこから生まれている？



# 機械学習モデルがなぜその予測をしたかの要因分析

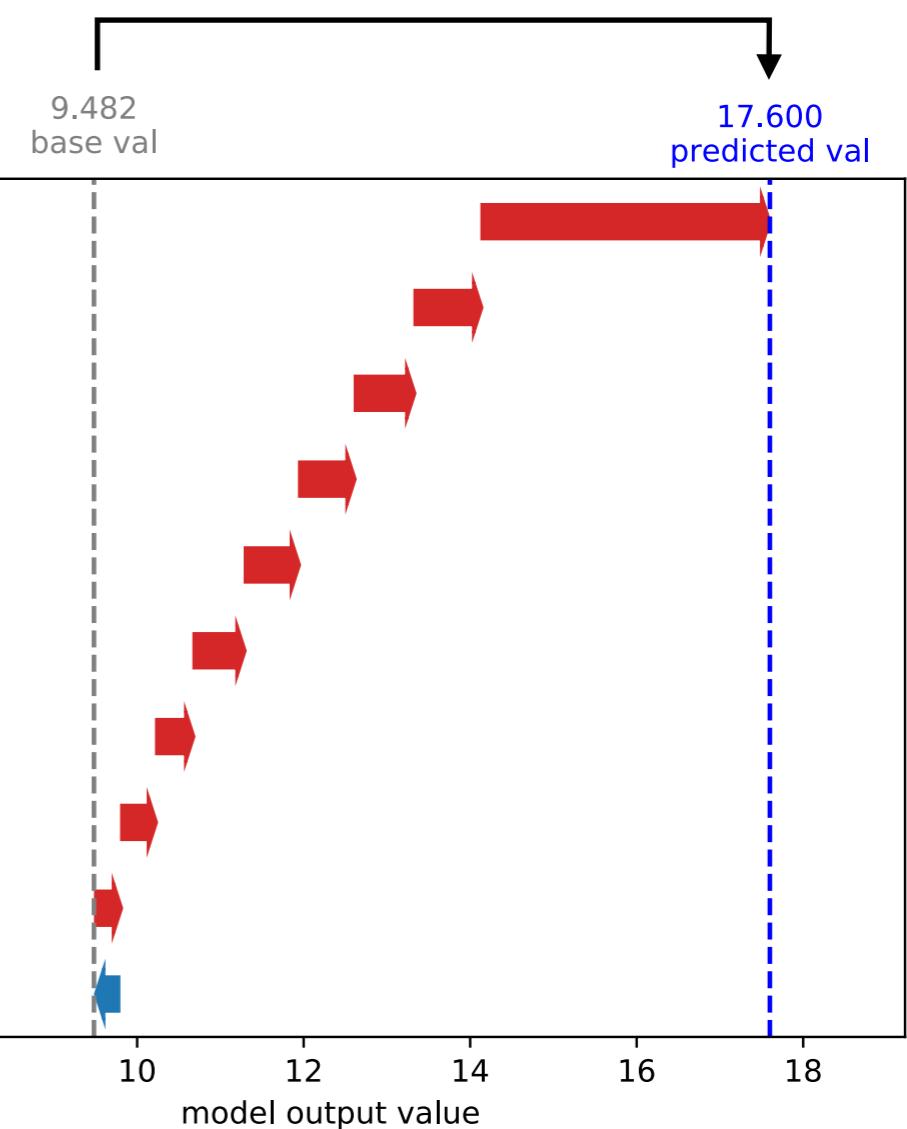
多人数の協力ゲームで得た報酬を各プレイヤへ公平に分配する  
ゲーム理論の問題とみなすことで各々の特徴量の寄与度を算出

|SHAP値|  
の降順

Composition: (1) Mg 83.46 (2) Li 16.53

**SHAP値の和 = データセット平均からの増加分**

	feat name	feat val	SHAP val
1	p(CH <sub>4</sub> )/p(O <sub>2</sub> )	2.0	3.459
2	electronegativity (2)	16.043	0.804
3	density (2)	8.832	0.718
4	delta fus H (1)	707.751	0.669
5	electronegativity (1)	102.657	0.653
6	delta fus H (2)	49.616	0.616
7	Impregnation	1.0	0.449
8	Therm.decomp.	0.0	0.419
9	density (1)	145.223	0.314
10	Temperature, K	1013.15	-0.281



# TreeExplainer: 決定木アンサンブル用のSHAP

一般には計算困難(NP困難)な量だが、決定木アンサンブルでは  
SHAP値が多項式求解可能 (TreeExplainer or treeSHAP)

インタラクティブな解析を提供するともこなれたツールもある  
→ <https://github.com/slundberg/shap>

ARTICLES

<https://doi.org/10.1038/s42256-019-0138-9>

nature  
machine intelligence

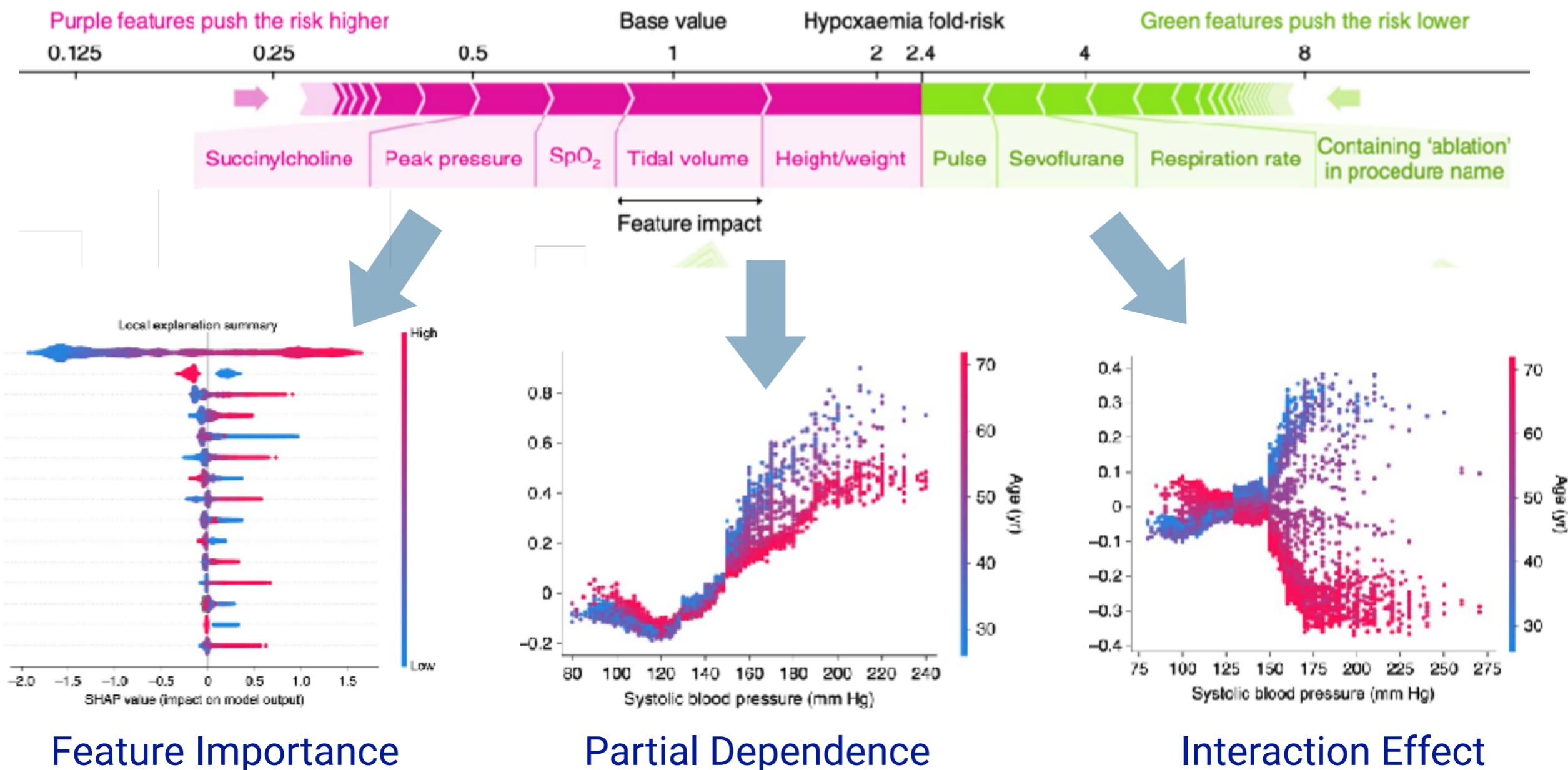
NATURE MACHINE INTELLIGENCE | VOL 2 | JANUARY 2020 | 56-67

## From local explanations to global understanding with explainable AI for trees

Scott M. Lundberg<sup>1,2</sup>, Gabriel Erion<sup>3</sup>, Hugh Chen<sup>2</sup>, Alex DeGrave<sup>2,3</sup>, Jordan M. Prutkin<sup>4</sup>, Bala Nair<sup>5,6</sup>,  
Ronit Katz<sup>7</sup>, Jonathan Himmelfarb<sup>7</sup>, Nisha Bansal<sup>7</sup> and Su-In Lee<sup>2\*</sup>

# SHAPによる学習済みモデルからの要因分析

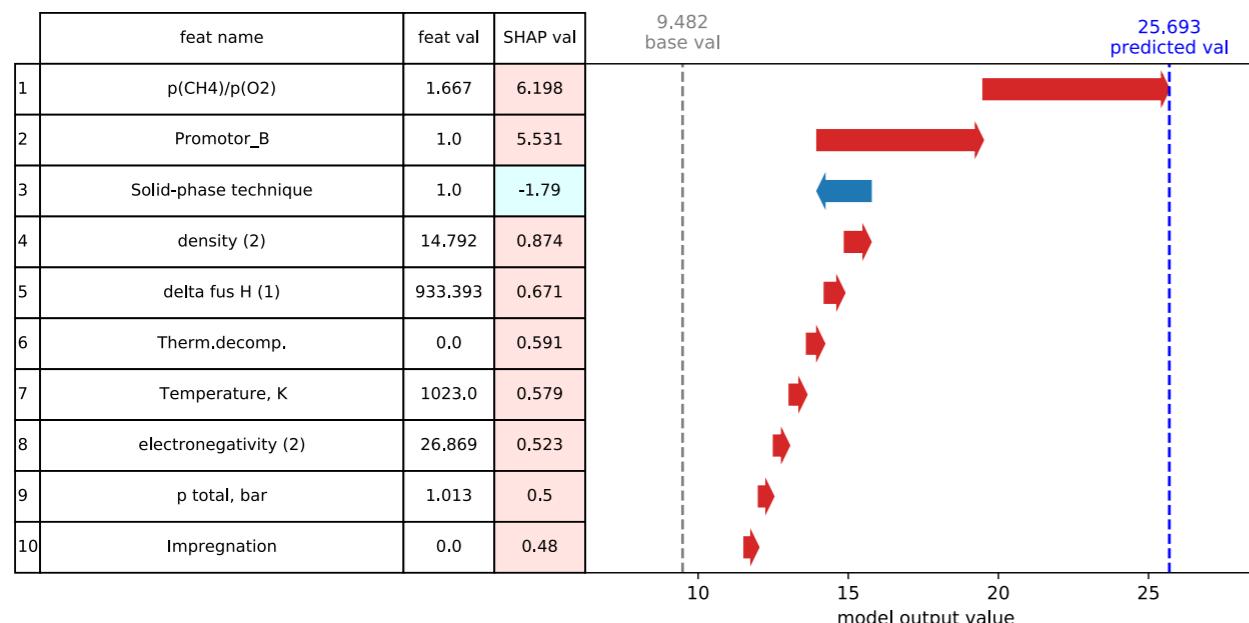
データや学習したモデルから得られる多角的情報を可視化などで抽出し、専門家と協働し専門知見や実制約に照らして利活用



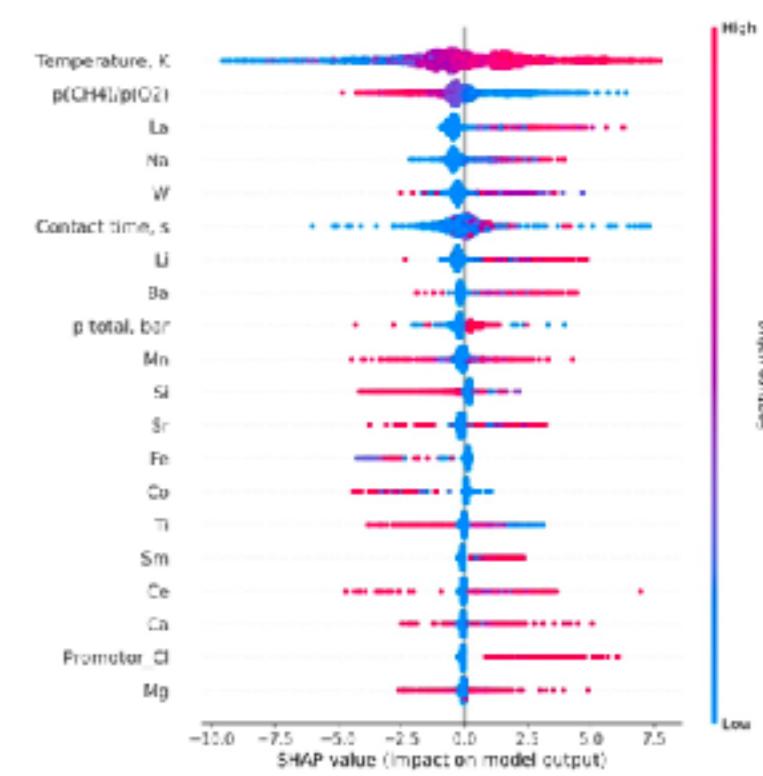
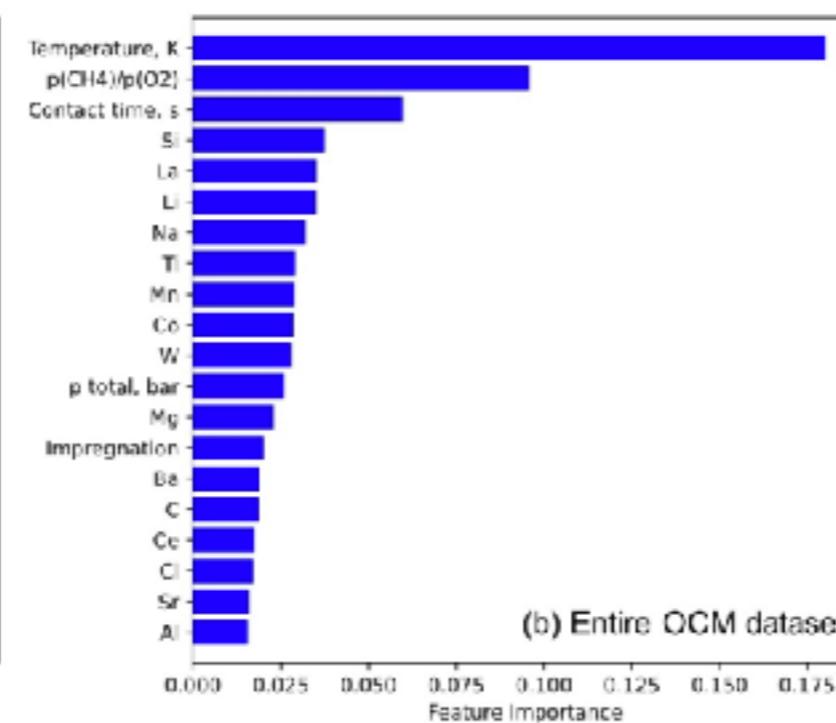
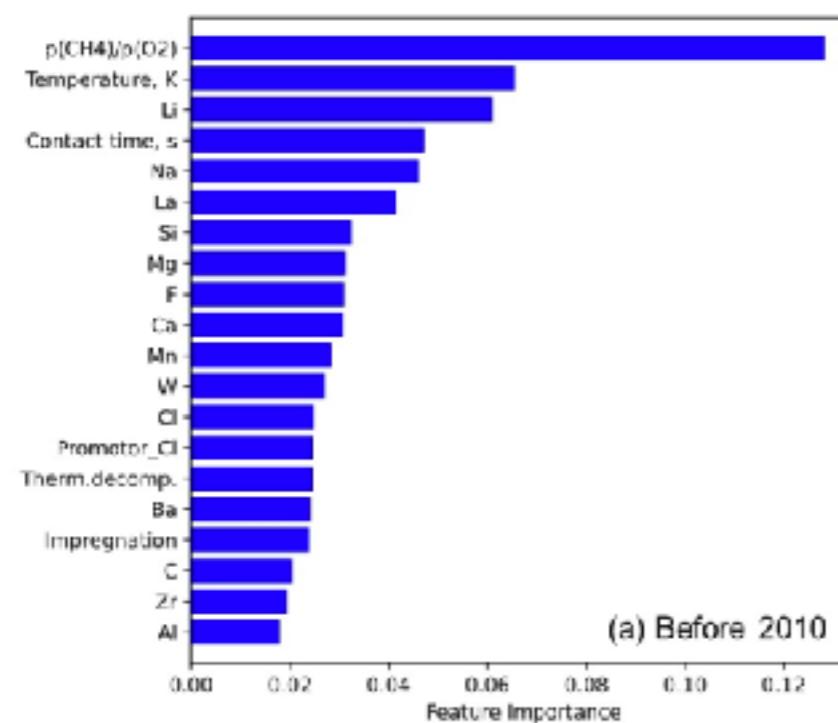
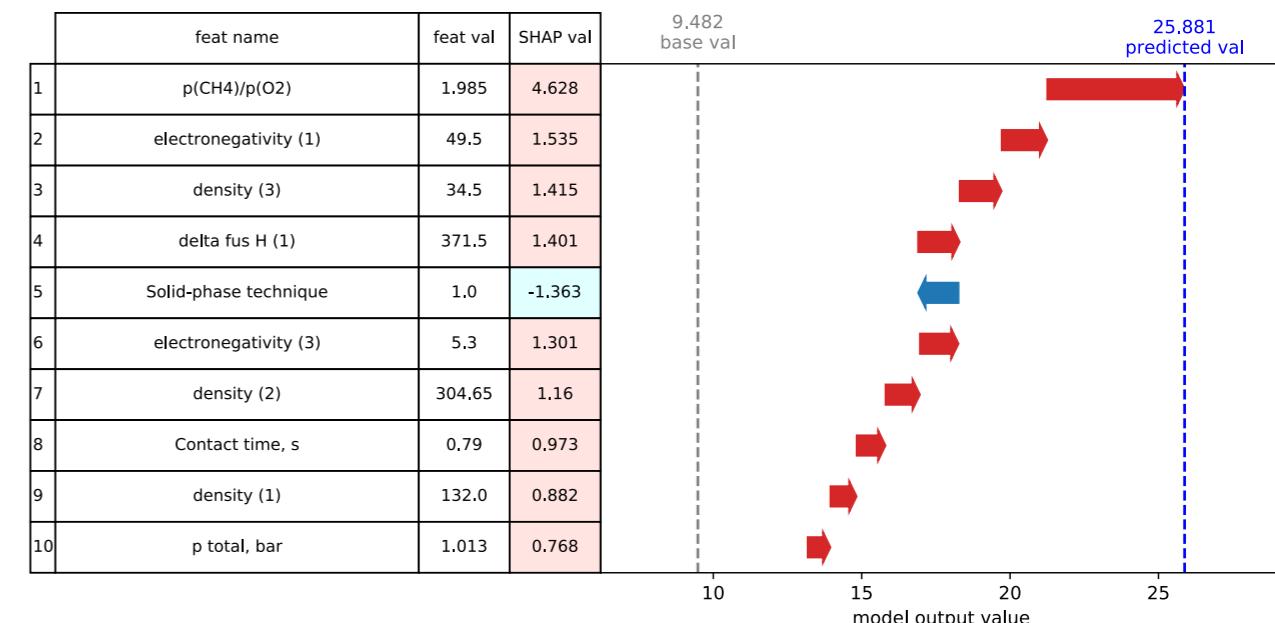
# SHAPによる学習済みモデルからの要因分析

Mine, S.; Takao, M.; Yamaguchi, T.; Toyao, T.\*; Maeno, Z.; Hakim Siddiki, S. M. A.; Takakusagi, S.; Shimizu, K.\*; Takigawa, I.\* *ChemCatChem* 2021. <https://doi.org/10.1002/cctc.202100495>.

1st: (1) Mn: 72.3 (2) Li: 27.7



2nd: (1) Sr:50.0 (2) Ce:45.0 (3) Yb:5.0



# 教訓

自然科学分野での利活用はMLの技術研磨だけでは成功しない。  
分野専門家との協働が必要不可欠

- MLがどういう技術なのか**MLの特性と限界**を正しく把握する
- 「データの収集計画(実験計画)と品質保証、適用範囲の理解」  
が“**“data-driven”**の心臓であることをいつも心に
- 「探索」が目的なら**MLの果たす役割はあくまで一部**と心得る
  - 👍 専門家との協働、分野の専門知識に照らした検証・解釈
  - 👍 シミュレーション・実験自動化・論理推論との融合

# 今日のテーマ

- **自己紹介 (機械学習と自然科学の境界)**
- **機械学習とは新しいプログラミングの方法**
- **機械学習屋は一体何が楽しいのか？**
  - 分子の表現と機械学習
  - グレイボックス最適化 (演繹 + 帰納)：論理学と統計学の融合？
- **自然科学研究で機械学習を使おうとすると必ずぶつかる本当に難しい問題**
  - データモデリングと予測アルゴリズム (The Two Cultures)
  - 予測か理解か：Rashomon効果, Underspecification, 解釈多様性
  - 人間の認知バイアスに由来する問題：仮説、失敗、成功バイアス、etc.
- **機械学習から機械発見へ**
  - 「発見」「理解」の道筋は合理化できるのか？自動化できるのか？

# 機械発見は機械学習の技術研磨だけでは到達しえない！

科学が求めること: 分からなかったことが分かる(科学的発見と科学的理解)

**理解**      原因と結果(因果関係)を見出す(そして客観的に説明できる)

$x \rightarrow y$  の過程を人間の限られた認知能力の範囲で理解する

**発見**      今まで見出されていない良い対象を見出す

$x \rightarrow y$  を利用して良い  $y$  を持つ  $x$  を**発見**する

- ✓ いざれも「機械学習」だけでは解けないことをまず理解すること。
- ✓ いざれも予測技術である機械学習のスコープ外の問題であり、機械学習以外のもの(介入実験、専門的知見、専門家との協働)が必須
- ✓ 「人工知能(AI)」と言うと既に解決技術があるような気がしてしまうが未解決問題。科学者と意識の食い違いが起こりやすく相互理解が重要
- ✓ **人工知能分野では「発見を自動化できるのか？」は重要な未解決課題！**  
→ 私たちは日々「発見」と「学習」を繰り返して生きている

# 科学的理解とは：自然科学研究は人間の営みである

## ✓ (自然のしくみを) 「人間が」 理解する必要がある

「理解」 に関する研究や技術が一筋縄ではいかないのはすべてこのせい

- **認知限界**：人間は大量の情報を理解できない・思い込み(仮説)が必須
  - **認知バイアス・社会バイアス**：人間がデータをとるとバイアスは不可避
  - **感情(心理)**：私たちの思考・判断は感情に訴える方法や情報操作に脆い
  - **科学的理解の成功バイアス**：科学的理解は出版されてはじめて世の中に広まり皆の科学的理解となるが、出版結果は真実の一部のみに強く偏る
  - また、自然の法則が必ずしも人間のしょぼい認知限界の範囲で簡潔に記述したり制御できるとは限らない
- 
- ✓ **手に入る情報は常に部分的**：人生が有限である以上、すべてをモデル化したり、ありとあらゆる情報を入力変数にしたりすることは不可能。この意味で何らかの「学習」や「知識」は不可避

# 実験自動化：誰だって単調で退屈な労働から解放されたい！

- ✓ 科学研究においても非効率な労働がいずれ自動化されるのは歴史的必然  
→ 科学につきまとう再現性・属人性・過酷労働の問題解消にも資する



Organic synthesis in a modular robotic system. *Science* 363 (2019)



A mobile robotic chemist. *Nature* 583 (2020)



Automating drug discovery. *Nature Reviews Drug Discovery* 17 (2018)



# ただし作業の自動化と発見そのものの自動化は別次元の問題

- ✓ ロボット系があれば綺麗な実験が(人手よりは)たくさんできるがそれでも「発見」には力技の加速化では到達できない(可能な候補空間が広すぎ)
- ✓ 自動化技術の研究と合わせて「何をどう発見するのか」の戦略が必須  
→ 探索する系・スコープ、機器で計測すべき変数、タスクのデザイン

**Machine Learning**

How to cite:

International Edition: [doi.org/10.1002/anie.201909987](https://doi.org/10.1002/anie.201909987)

German Edition: [doi.org/10.1002/ange.201909987](https://doi.org/10.1002/ange.201909987)

## Autonomous Discovery in the Chemical Sciences Part I: Progress

Connor W. Coley,\* Natalie S. Eyke, and Klavs F. Jensen\*

**Computer Chemistry**

How to cite:

International Edition: [doi.org/10.1002/anie.201909989](https://doi.org/10.1002/anie.201909989)

German Edition: [doi.org/10.1002/ange.201909989](https://doi.org/10.1002/ange.201909989)

## Autonomous Discovery in the Chemical Sciences Part II: Outlook

Connor W. Coley,\* Natalie S. Eyke, and Klavs F. Jensen\*



# 人工知能分野における「発見」の研究

**KAKEN**

研究課題をさがす

研究者をさがす

KAKENの使い方

日本語 ▾

## 巨大学術社会情報からの知識発見に関する基礎研究

研究課題

研究課題/領域番号 10143106

サマリー ▾

研究種目 特定領域研究(A)

配分区分 補助金

研究機関 九州大学

研究代表者 有川 節夫 九州大学, 大学院・システム情報科学研究院, 教授 (40037221)

研究分担者 丸岡 章 東北大学, 大学院・情報科学研究科, 教授 (50005427)

佐藤 泰介 東京工業大学, 大学院・情報理工学研究科, 教授 (90272690)

佐藤 雅彦 京都大学, 大学院・情報学研究科, 教授 (20027307)

金田 康正 東京大学, 情報基盤センター, 教授 (90115551)

宮野 悟 東京大学, 医科学研究所, 教授 (50126104)

研究期間 (年度) 1998 – 2000

研究課題ステータス 完了 (2001年度)

配分額 ・注記 66,400千円 (直接経費: 66,400千円)

2001年度: 3,000千円 (直接経費: 3,000千円)

2000年度: 15,300千円 (直接経費: 15,300千円)

1999年度: 19,100千円 (直接経費: 19,100千円)

1998年度: 29,000千円 (直接経費: 29,000千円)

キーワード

発見科学 / 知識科学 / データマイニング / データベース / 科学的発見の論理 / アブダクション / 機械学習 / ネットワークエージェント / 知識発見



→ 発見科学、科学的発見の論理、知識発見、機械学習、アブダクション

# 発見科学と機械発見



## 機械学習から機械発見へ

Our Studies on Machine Learning and Machine Discovery

有川 節夫\*  
Setsuo Arikawa

\* 九州大学大学院システム情報科学研究科情報理学専攻  
Dept. of Informatics, Kyushu University.

1996年8月28日受理

**Keywords:** machine learning, machine discovery, algorithmic learning theory, discovery science.

### 1. はじめに —創造工学を機械化できないか—

1968年から1972年の頃であったと思う。北川敏先生や国沢清典先生、森口繁一先生達の企画で日本科学技術連盟主催のセミナー(講習会)が定期的に開催されていた。その一つに「創造工学」があった。KJ法や等価変換論といった代表的な創造工学の手法の提唱者自身による講演があり、私も、北川先生の好意で出席させてもらった。その当時計算理論の研究をしていたので、このような講演が非常に新鮮に感じられ、また、そうした創造工学の手法が、非常に主観的で精神論的なものに感じられた。

もう少し客観的に機械的にそうした手法を実現できないものか、そのような研究はきっと非常に重要になり盛んになるはずだ、というような生意気なことを北川先生に話したように記憶している。

また、1970年代には日米計算機会議というのが、2~3年おきに開催され、その最初の会議で、アメリカのA.W.Biermannが、有限オートマトンを対象にした、文法推論に関する非常に興味深い研究を発表し

それを記述するプログラムがデータのサイズそのものとほとんど変わらないとき、ランダムであるという。したがって、データにアルゴリズム的な規則性がなければ、ランダムということになり、文法推論可能性や学習可能性と対極をなす概念と考えられる。文法推論可能であれば、データ圧縮が可能であるという観点から、簡単な報告を書き、この方面的研究を本格的に展開するつもりでいた。

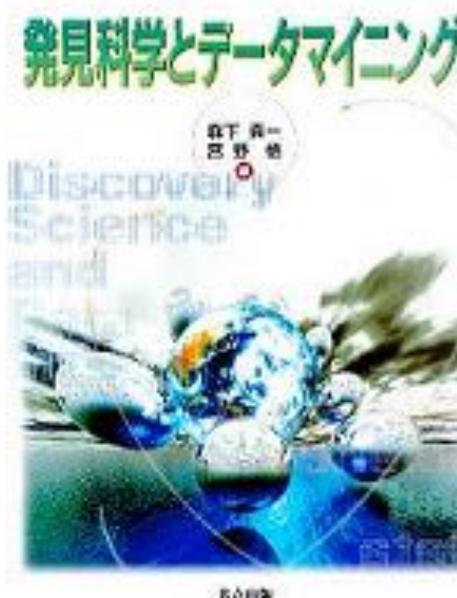
### 2. 情報検索は人工知能の基礎である

しかし、その頃スタートしたデータベースと情報検索システムに関連した特定研究で、北川先生が責任者の一人であったため、情報検索関連の研究をすることになり、この計画は無期延期になった。情報検索の研究では、学術情報の生産者でありかつユーザーである研究者をシステムのなかに積極的に位置づけ、研究者の主觀や偏見を検索に反映できる知的なシステムを構築した。これは、研究者が使い込んでいくとどんどん賢くなり、自分自身の知識が検索に生かされるようになるもので、当時としては非常に斬新なシステムとして、JICSTなどでも評価していただいた。情報検索

# 発見科学と機械発見

## 発見科学とデータマイニング

この書籍は現在お取り扱いできません。



森下 真一・宮野 悟 編

ISBN 978-4-320-12018-1

判型 A4変型

ページ数 318ページ

発行年月 2001年06月

価格 4,840円（税込）

### 序章 日本の発見科学プロジェクト（有川節夫）

#### 第I部 推論による知識発見

##### 第1章 「発見」の科学哲学—歴史的素描（野家啓一）

第2章 統計的記号処理言語PRISM（亀谷由隆・佐藤泰介）

第3章 予測モデルからのルール抽出—数式から言語へ（月本 洋・森田千絵）

第4章 帰納論理プログラミングと証明補完（山本章博・有村博紀・平田耕一）

第5章 KeyGraph—キーワード抽出ツールから発見ツールへの展開（砂山 渡・大澤幸生・谷内田正彦）

第6章 IntelligentPadの合成と再利用—帰納推論の立場から（原口 誠・平田 淳）

#### 第II部 計算学習理論に基づく知識発見

##### 第7章 能動学習と発見科学（安倍直樹・馬見塚 拓）

第8章 くり返しゲームとしての学習アルゴリズム（丸岡 章・瀧本英二）

第9章 コンピュータサイエンスのための単純かつ効率的なサンプリング技法（渡辺 治）

第10章 学習アルゴリズムの評価（上原邦昭）

第11章 幾何クラスタリングの情報計算幾何構造（今井 浩）

第12章 Support Vector Machineによる分類（高須淳宏）

Pat LangleyとHeikki Mannilaの  
重要論文の和訳もついている

#### 第III部 機械学習とデータマイニングに基づく知識発見

##### 第13章 コンピュータ支援による科学的知識の発見（Pat Langley著／宮野 悟・丸山 修訳）

第14章 学習か、マイニングか、モデリングか？—古生態学からの事例研究（Heikki Mannila et al.著／森下真一訳）

第15章 分枝限定法を用いた並列グラフ探索による最適結合ルールの発見（中谷明弘・森下真一）

##### 第16章 知識発見と自己組織型の統計モデル（北川源四郎・樋口知之）

第17章 顧客の購買履歴からのデータマイニング（矢田勝俊・加藤直樹・羽室行信）

第18章 発見システムとヒューマンエキスペートのインテグレーション（丸山 修・宮野 悟）

#### 第IV部 大規模数値データからの知識発見

第19章 太陽地球系物理学への知識発見の応用（家森俊彦・上野玄太・能勢正仁・町田 忍・荒木 徹・亀井豊永・竹田雅彦）

第20章 ブラインドセパレーションとウェーブレットによる隠蔽画像の発見（新島耕一）

##### 第21章 計算機による科学的法則・モデルの発見方法の展開（鷲尾 隆・元田 浩）

第22章 多変量データからの多項式型法則の発見（中野良平・斎藤和巳）

第23章 音声データベースからの音声知識の発見（鈴木基之・牧野正三）

第24章 仮想化された人体からのナビゲーションに基づく知識発見の支援ツール（齋藤豊文・鳥脇純一郎）

#### 第V部 ネットワーク環境における知識発見

第25章 ミームメディアを用いた知財流通と科学技術データの可視化（田中 譲）

第26章 ズーミング技術を用いた対話的情報検索インターフェース（豊田正史・柴山悦哉）

第27章 リンク情報からの知識網構成（廣川佐千男・池田大輔・田口剛史）

第28章 インターネットでの企業間情報共有に向けたマルチエージェントシステム（毛利隆夫・高田裕志）

# 表現と介入：機械学習から機械発見へ至るための指針



## 表現: 必要十分な情報を機械に入れる

- ✓ 適切な入力表現のデザインと学習  
分子表現とGNNs、特に事前学習とその転移  
入力特徴量(記述子)の設計・エンジニアリング
- ✓ 機械学習モデル(入出力マッピング)の表現  
理論計算+機械学習(グレイボックス最適化)

## 介入: 相関から因果の根拠を得る

- ✓ 相関を因果と確証するための介入実験研究  
探索ポリシーと適応的実験計画のデザイン
- ✓ シミュレーションや実験自動化との融合

# 機械学習・機械発見にとっても実世界検証のための梁山泊



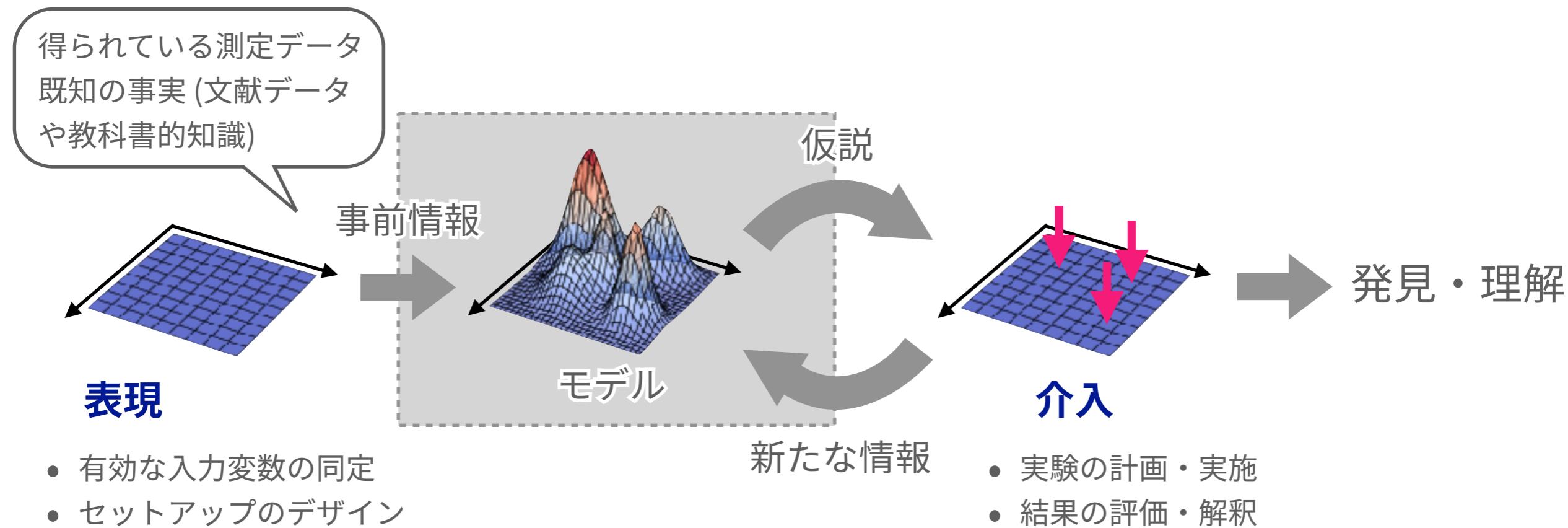
An exciting “real-world” test bench for ML researchers!

## ✓ 機械学習 (Machine Learning)

離散構造/組合せ構造を伴う機械学習 (分子、反応、反応経路Network)

## ✓ 機械発見 (Machine Discovery)

理論計算+ML (グレイボックス最適化)、実験+ML (探索と介入計画)

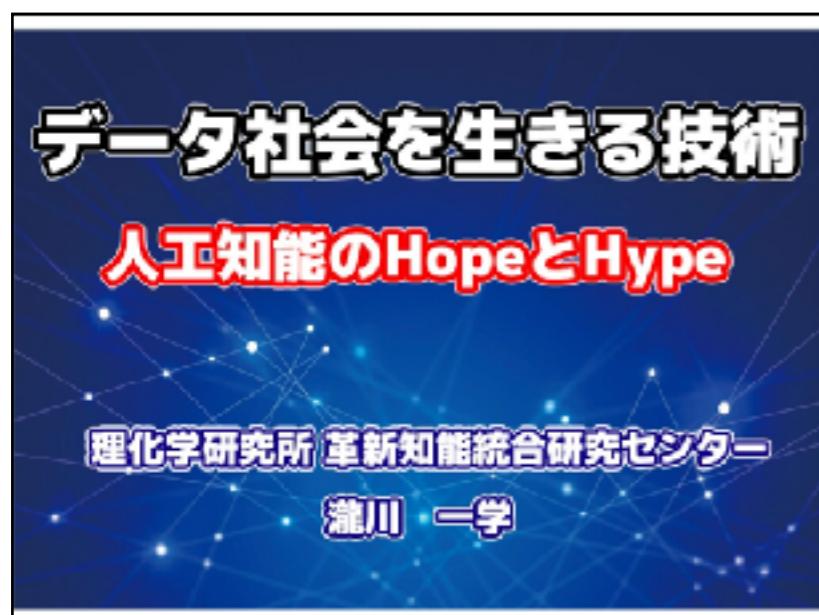
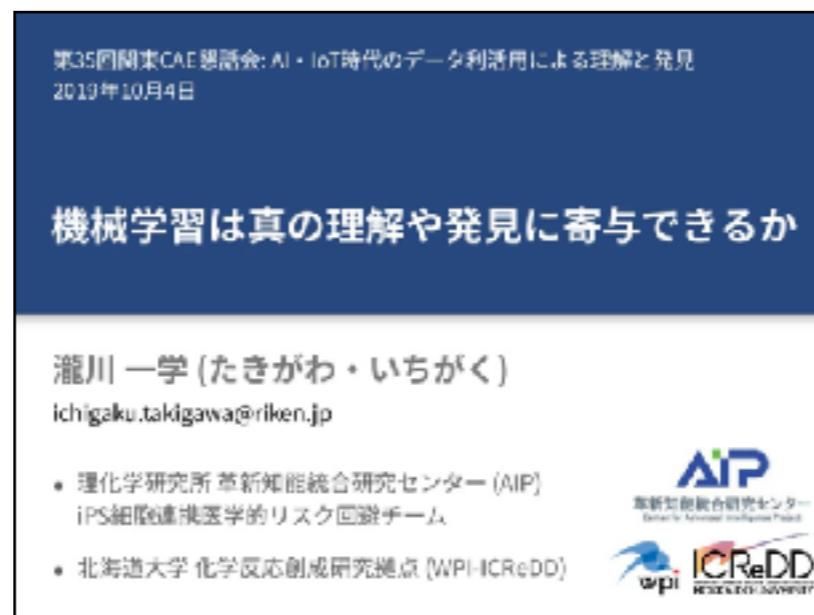


# このスライドも置いてあります

瀧川一学



<https://itakigawa.github.io/news.html>  
<https://www.slideshare.net/itakigawa/presentations>



# まとめ：今日の話

[https://itakigawa.github.io/data/talk\\_20211026.pdf](https://itakigawa.github.io/data/talk_20211026.pdf)

- **自己紹介 (機械学習と自然科学の境界)**
- **機械学習とは新しいプログラミングの方法**
- **機械学習屋は一体何が楽しいのか？**
  - 分子の表現と機械学習
  - グレイボックス最適化 (演繹 + 帰納) : 論理学と統計学の融合?
- **自然科学研究で機械学習を使おうとすると必ずぶつかる本当に難しい問題**
  - データモデリングと予測アルゴリズム (The Two Cultures)
  - 予測か理解か : Rashomon効果, Underspecification, 解釈多様性
  - 人間の認知バイアスに由来する問題 : 仮説、失敗、成功バイアス、etc.
- **機械学習から機械発見へ**
  - 「発見」「理解」の道筋は合理化できるのか？自動化できるのか？