

機械学習と機械発見

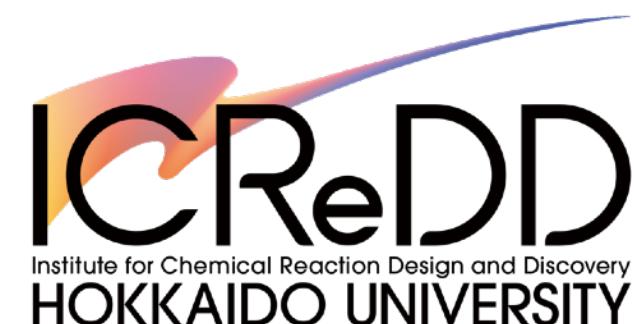
データ中心型の化学・材料科学の教訓とこれから

2021年11月12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
北海道大学 化学反応創成研究拠点 (ICReDD)





たきがわ いちがく
瀧川 一学

<https://itakigawa.github.io>

機械学習を研究している技術屋デス！

- 香川県高松市生まれ
- 1995～2004 北海道大 (工学研究科)
2004 博士(工学) "劣決定信号源分離の解の理論分析"
- 2005～2011 京都大 (化学研究所/薬学研究科)
バイオインフォマティクスセンター 助教
- 2012～2018 北海道大 (情報科学研究科)
大規模知識処理研究室 准教授
2015～2018 JSTさきがけ (材料インフォマティクス)
- 2019～ 北海道大学 化学反応創成研究拠点(ICReDD)
2019～ 理化学研究所 革新知能統合研究センター(AIP)

普段は京都大iPS細胞研との連携ラボ@京阪奈にいます
(iPS細胞連携医学的リスク回避チーム)

今日の話：機械学習の力の光明面と暗黒面

機械学習と機械発見の技術屋から見たデータ中心型の化学・材料科学の教訓とこれから

ライトサイド（光明面）

機械学習は「データを予測に変える」強力なテクノロジー！

- 分子の表現学習とGraph Neural Networks
- 帰納バイアスの設計とグレイボックス最適化

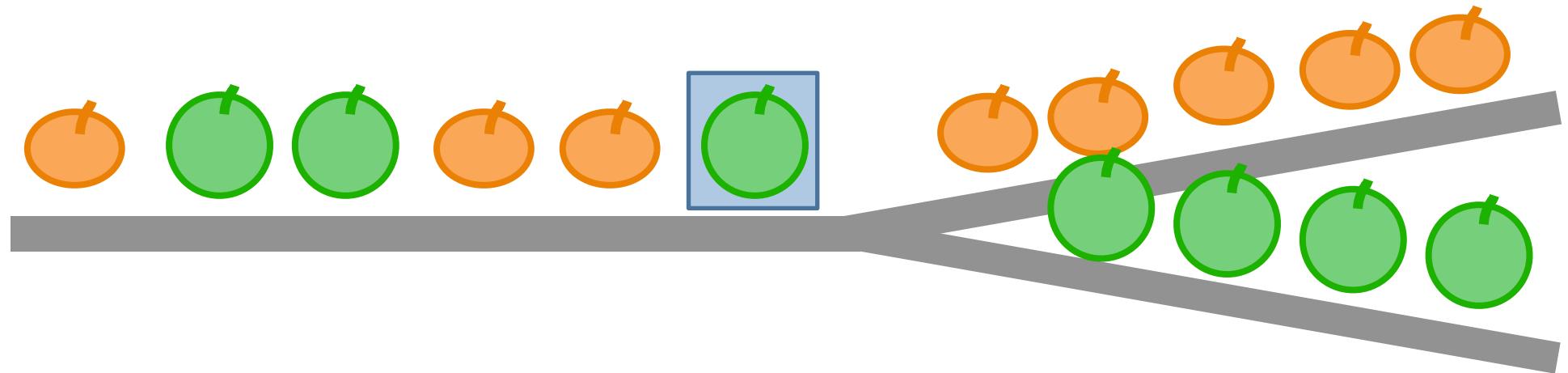
ダークサイド（暗黒面）

自然科学の実現象データで使うのはいろいろ激ムズ！！！

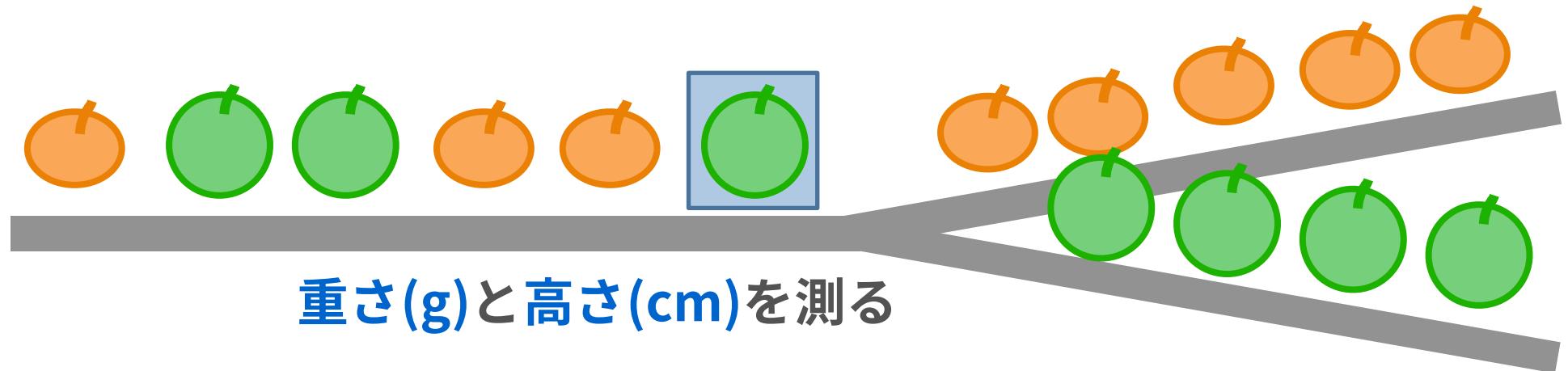
- 羅生門効果とUnderspecification
- 「予測ができること」は「理解」や「発見」ができるることを意味しない！

May the ML force be with you...

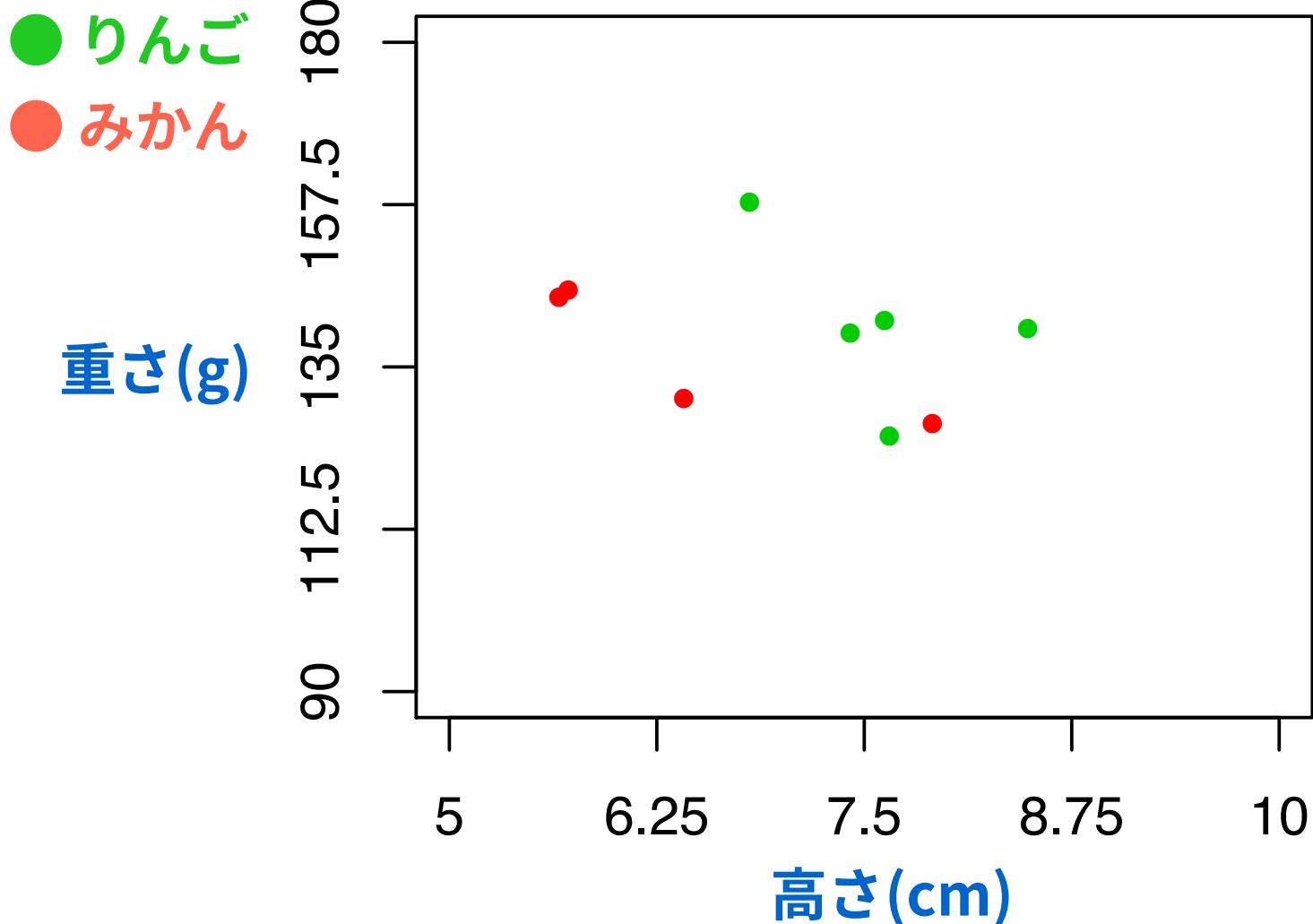
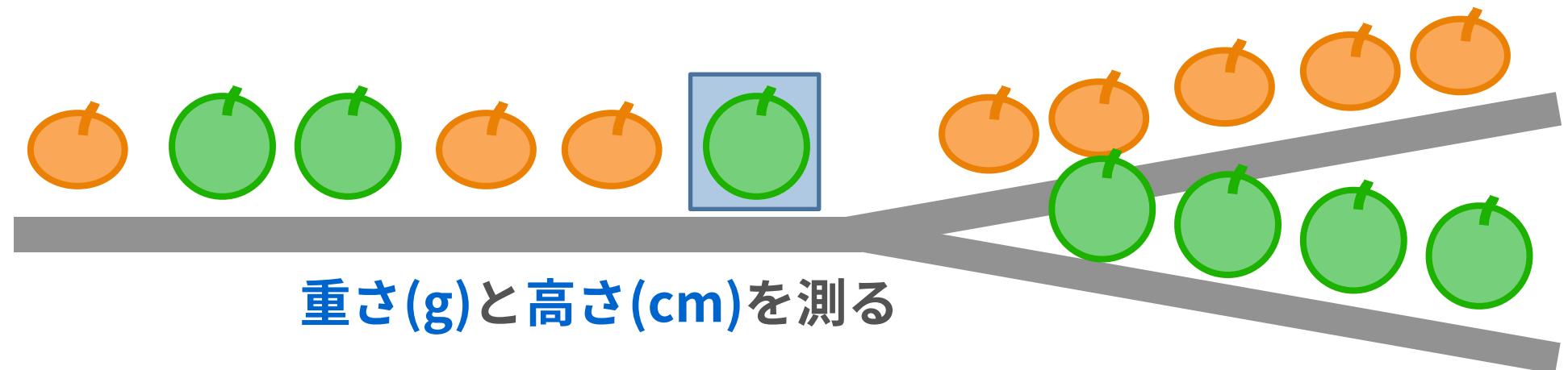
機械学習は「データを予測に変える」



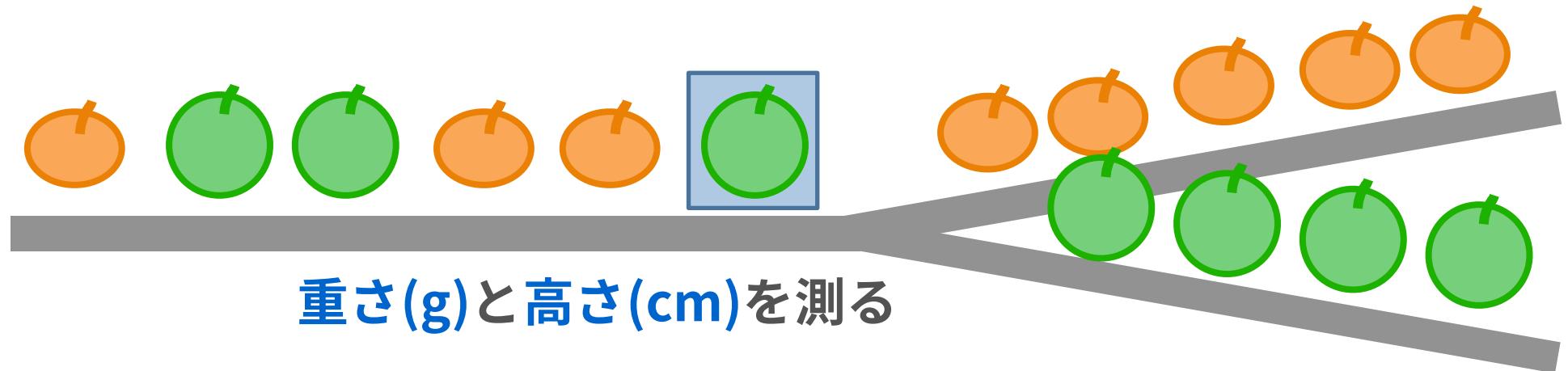
機械学習は「データを予測に変える」



機械学習は「データを予測に変える」

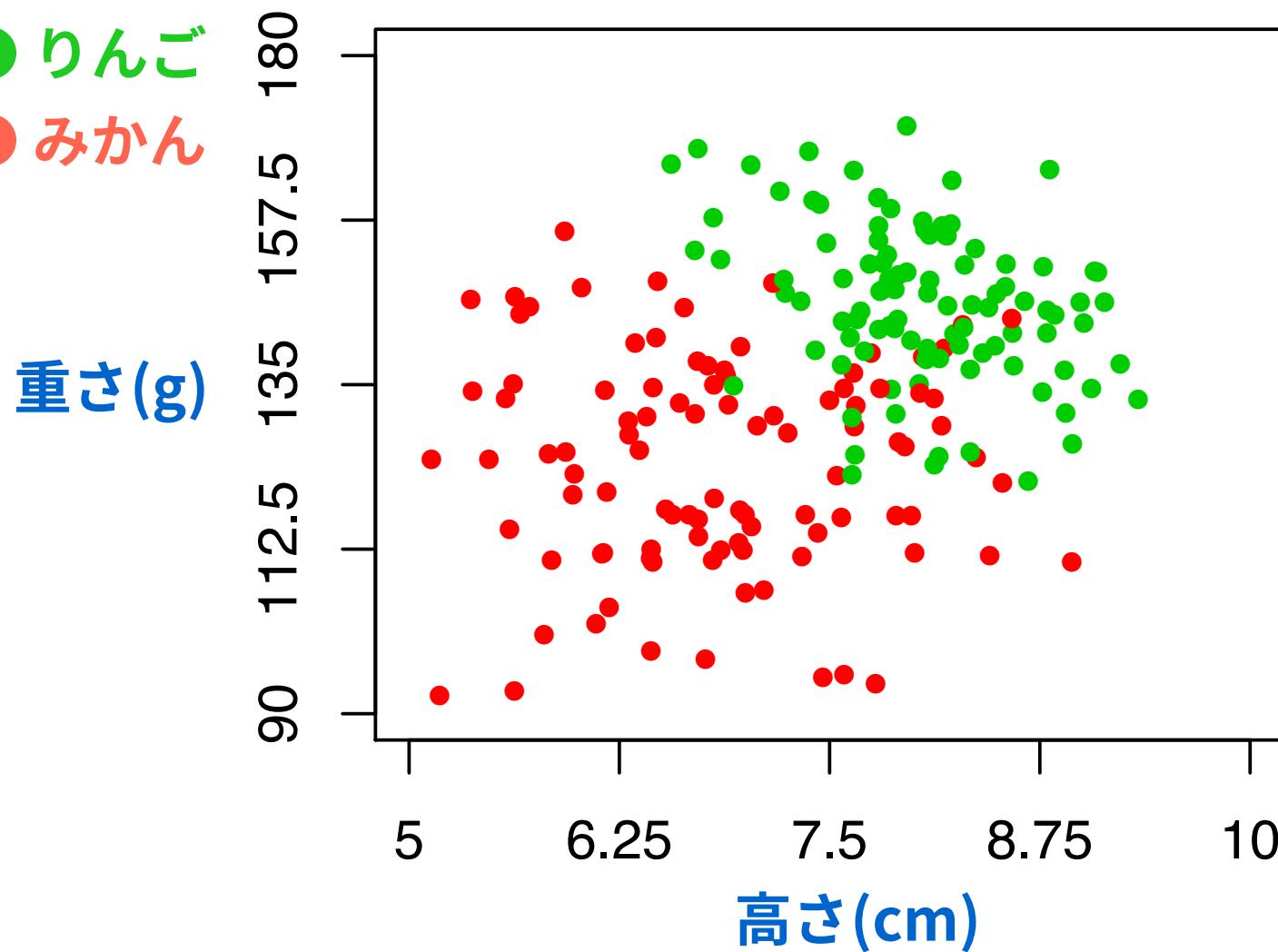


機械学習は「データを予測に変える」

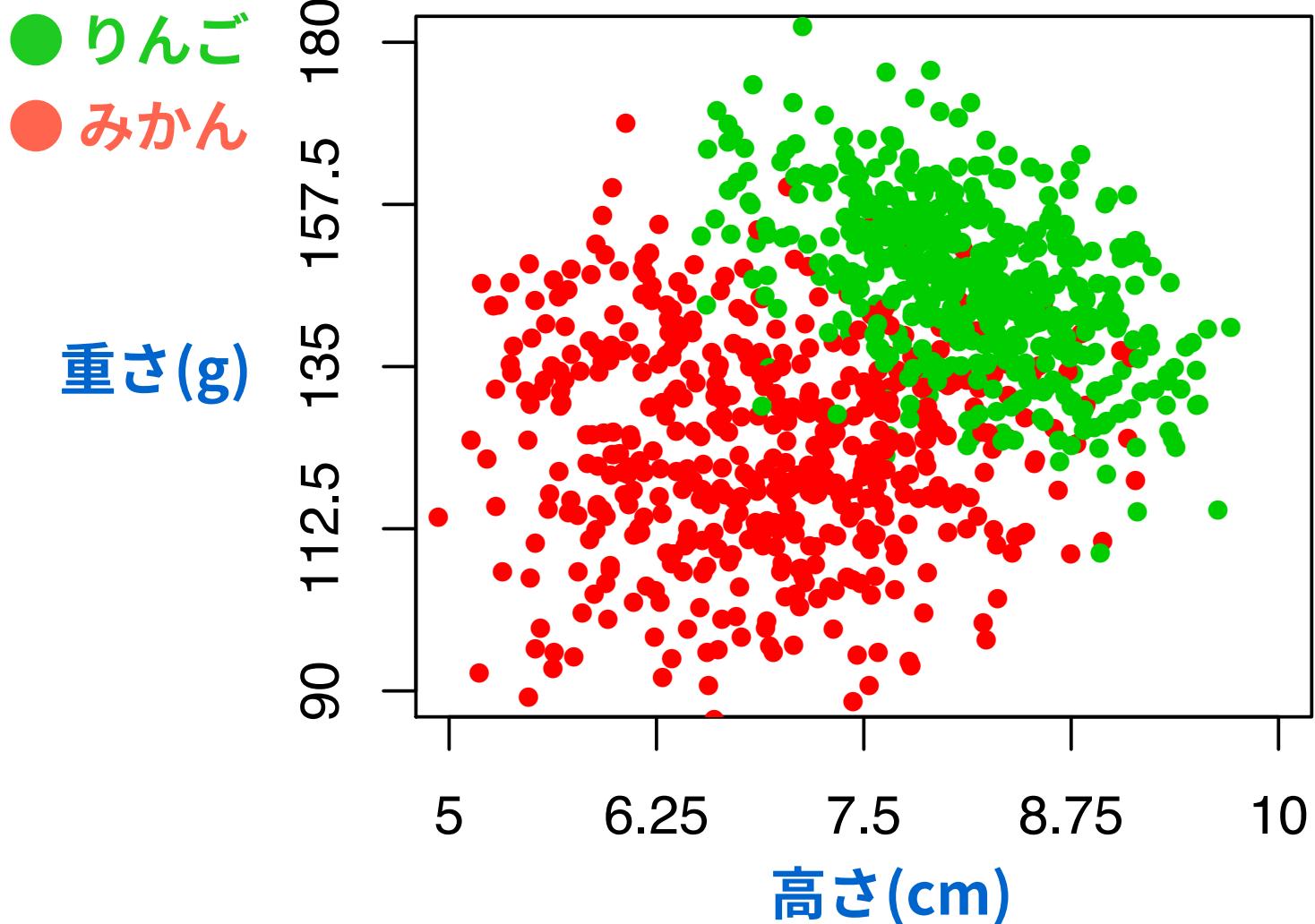
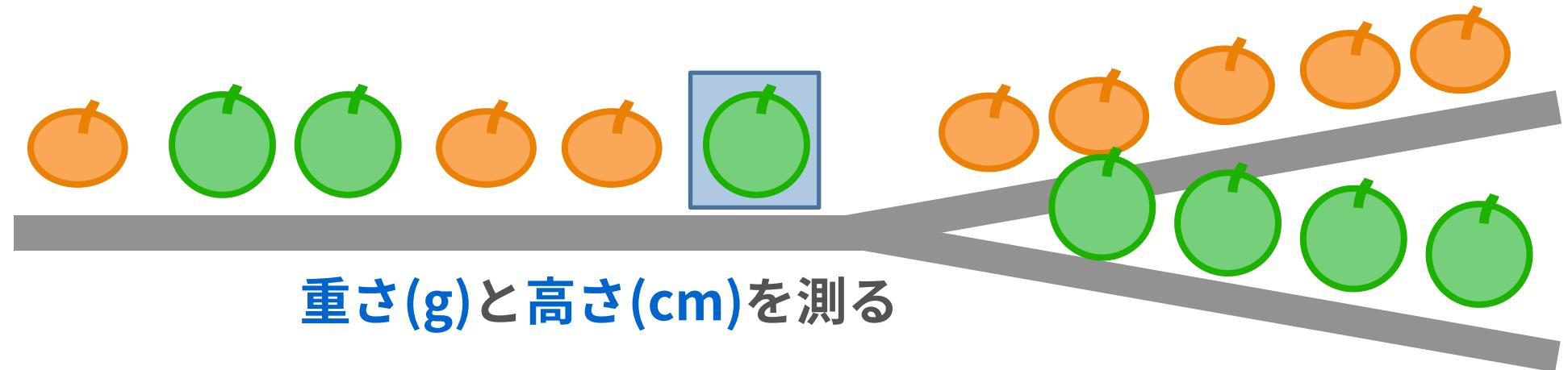


重さ(g)と高さ(cm)を測る

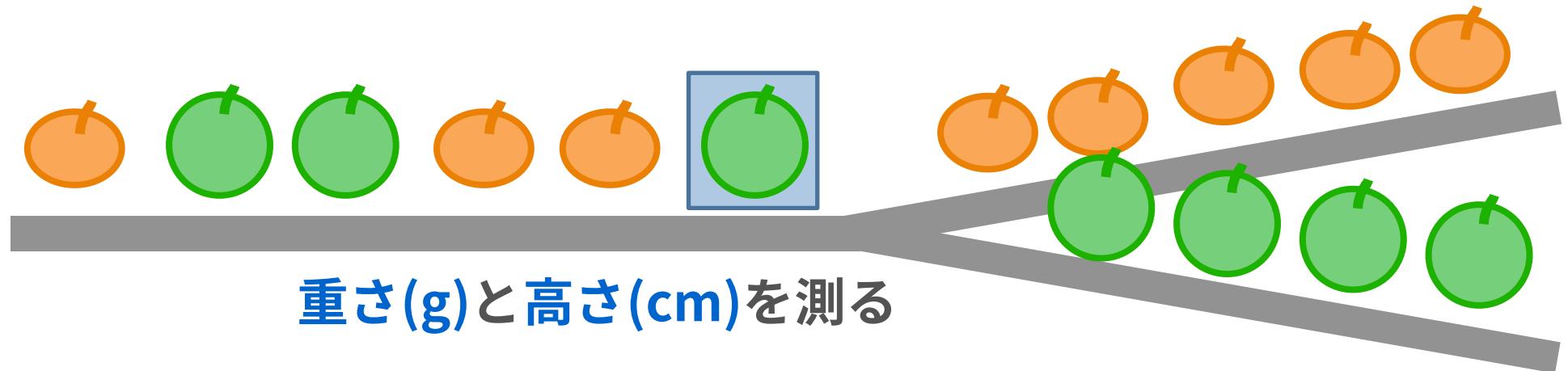
りんご
みかん



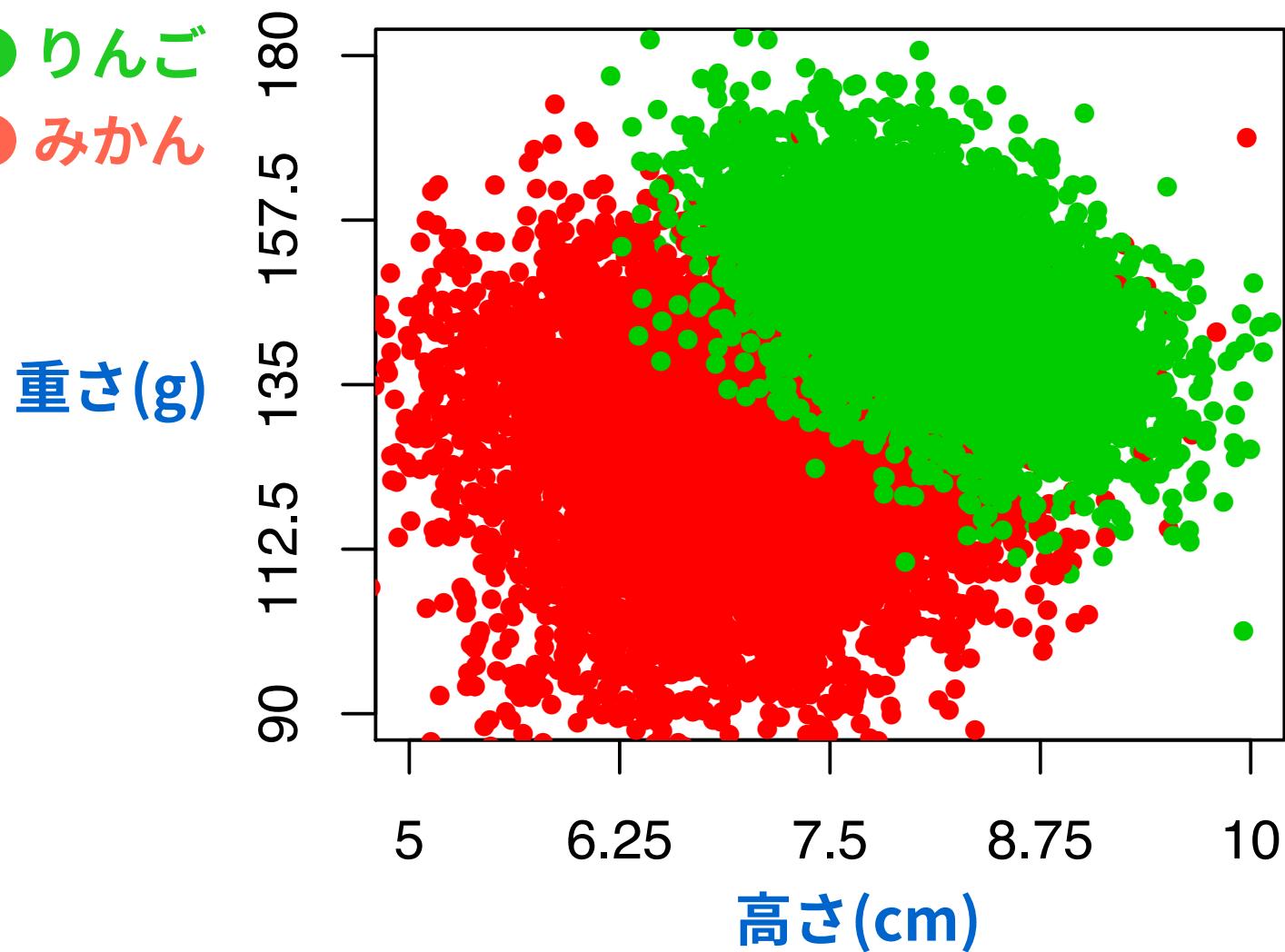
機械学習は「データを予測に変える」



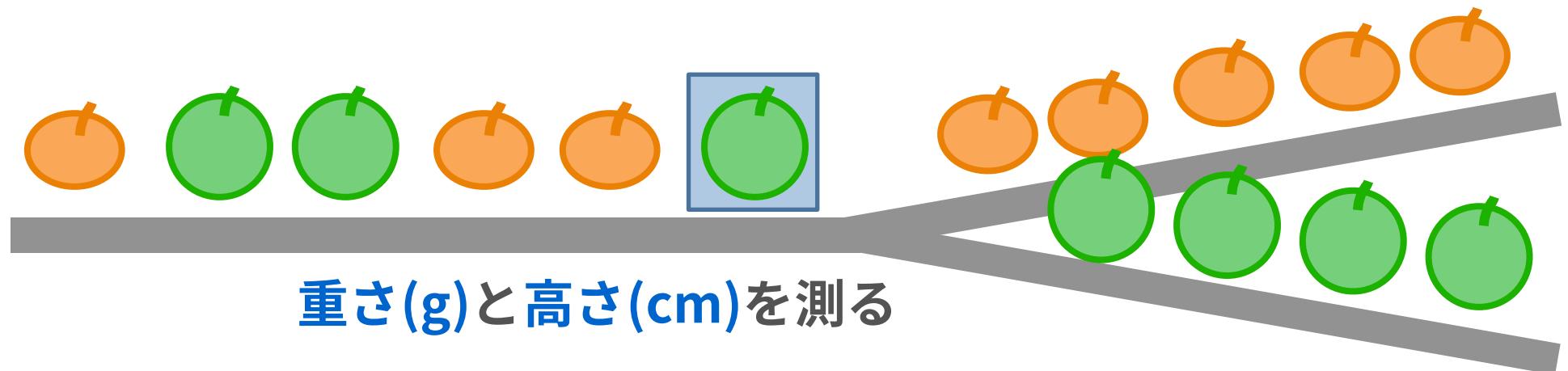
機械学習は「データを予測に変える」



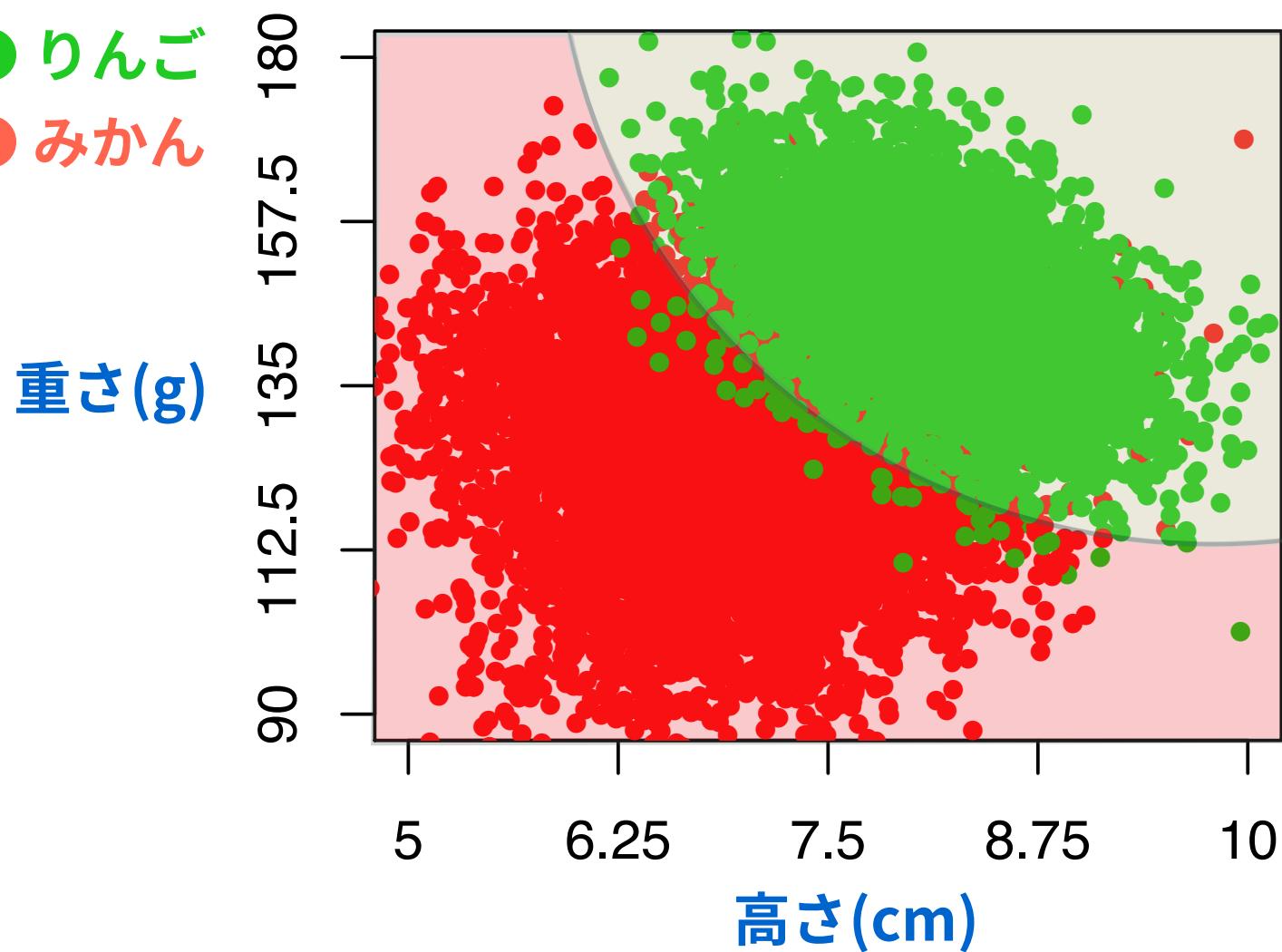
りんご
みかん



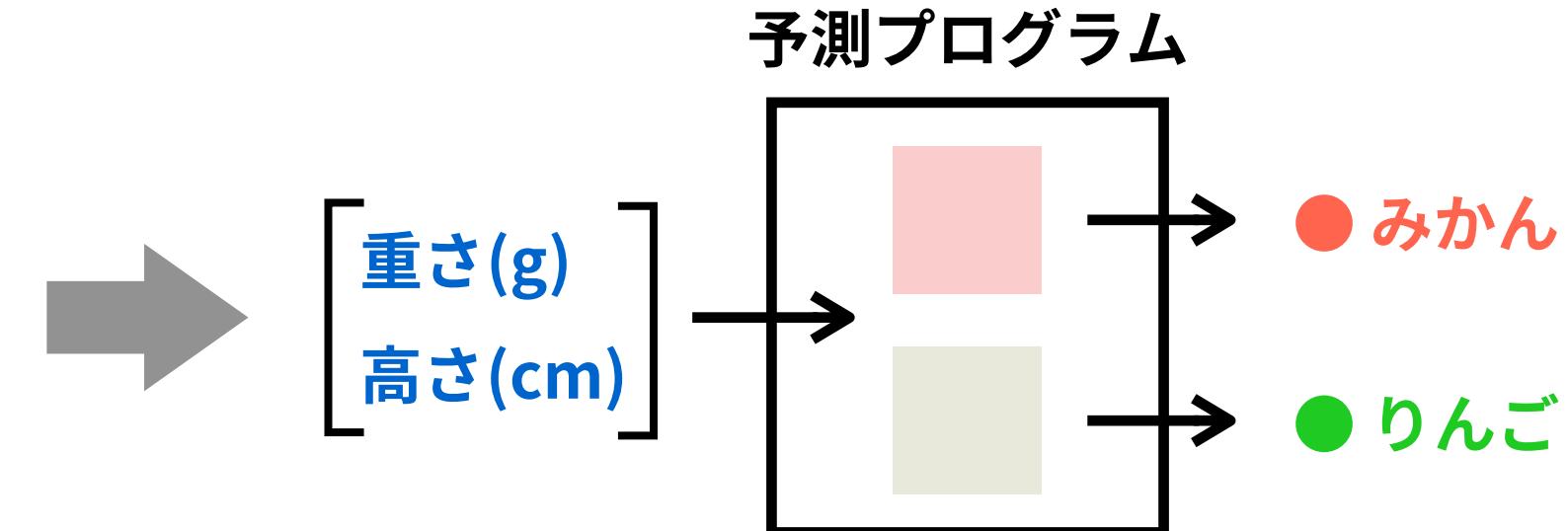
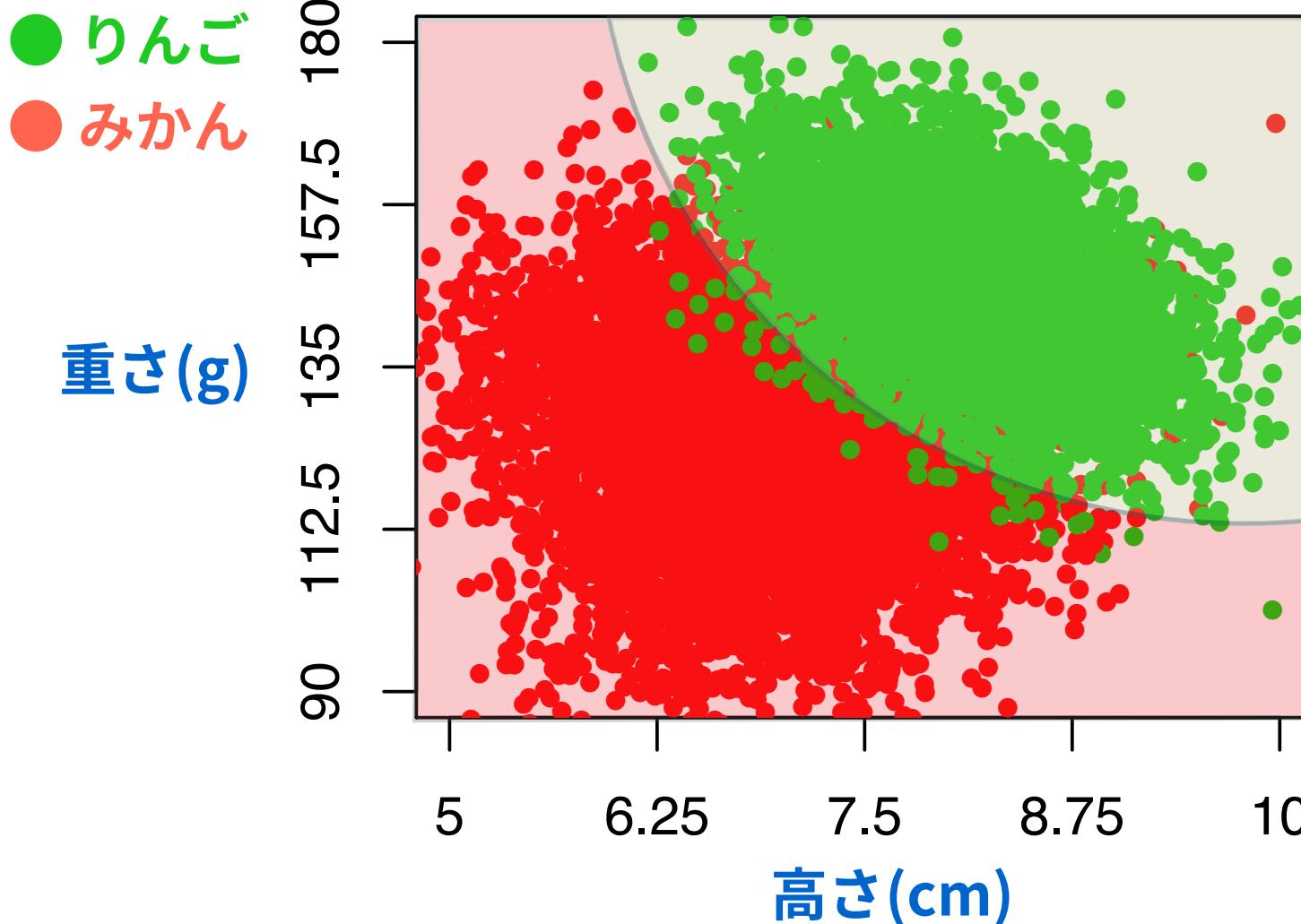
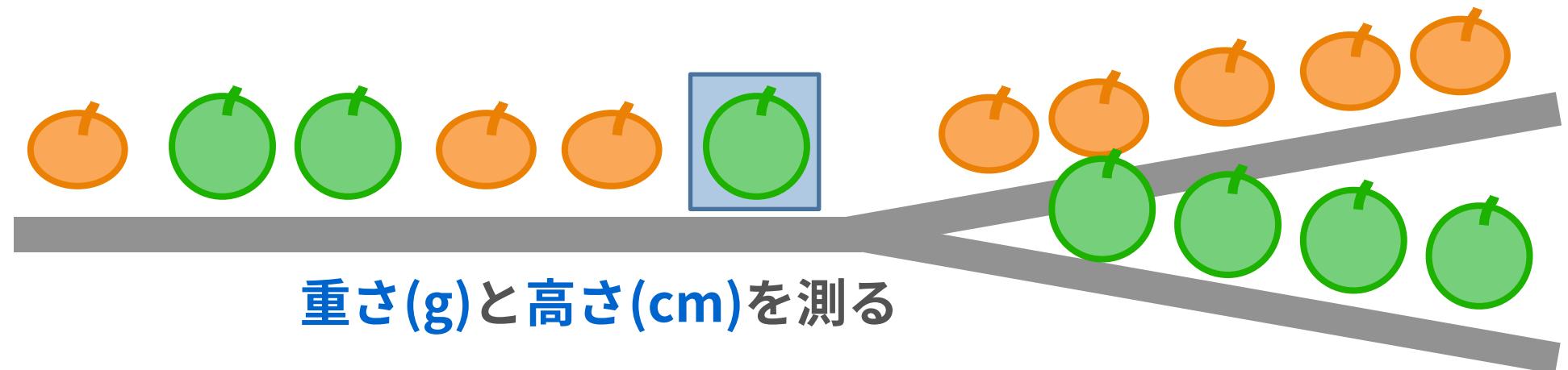
機械学習は「データを予測に変える」



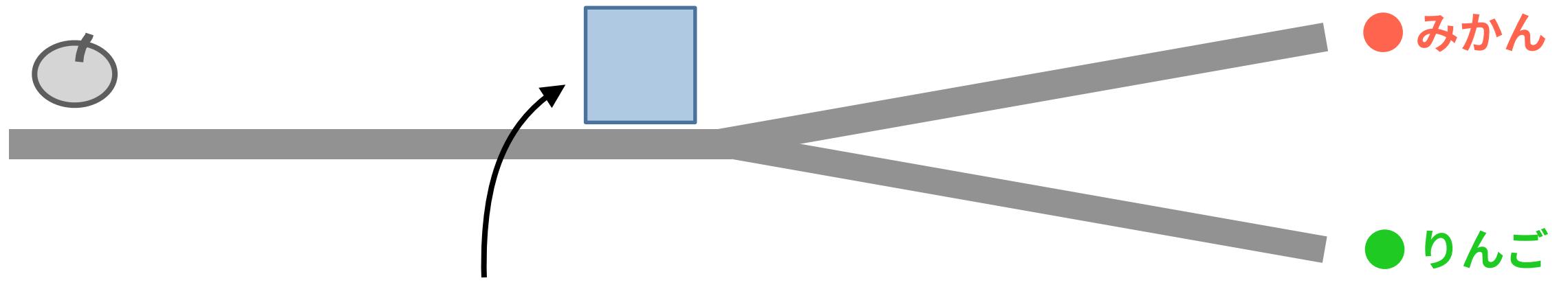
りんご
みかん



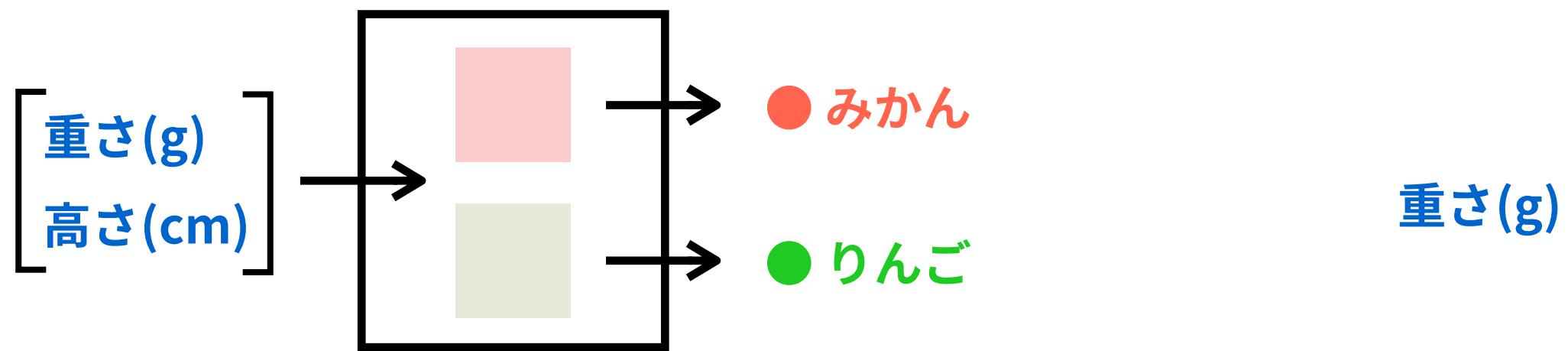
機械学習は「データを予測に変える」



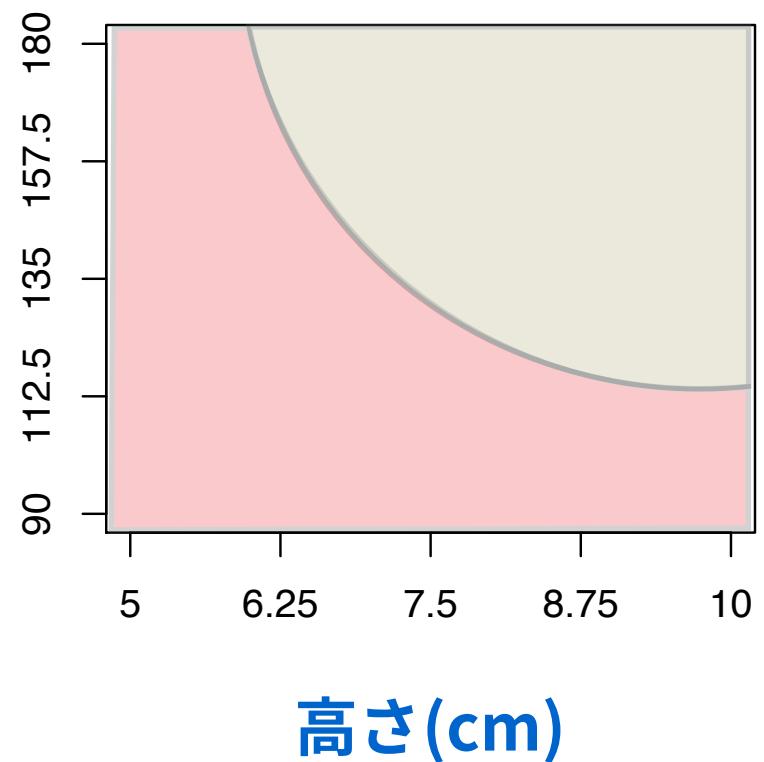
機械学習は「データを予測に変える」



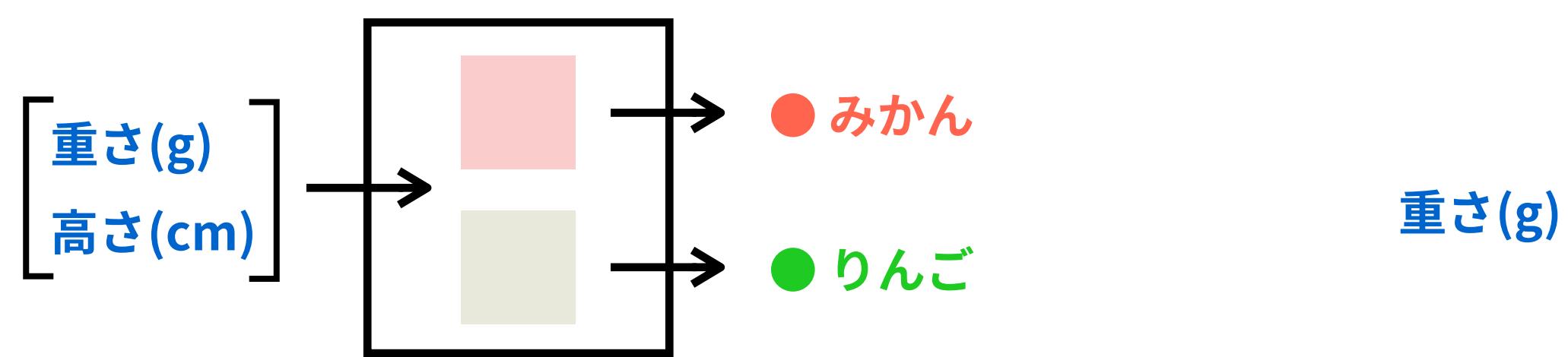
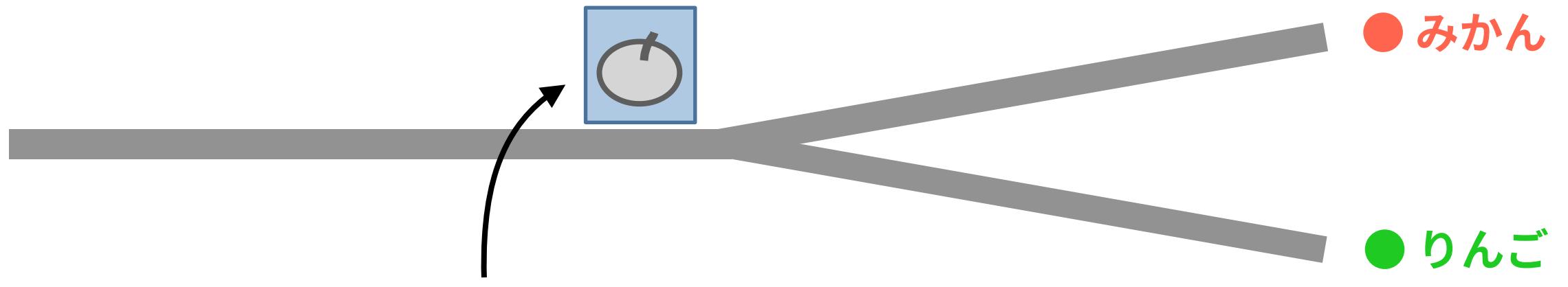
見本データから作っておいた予測プログラム



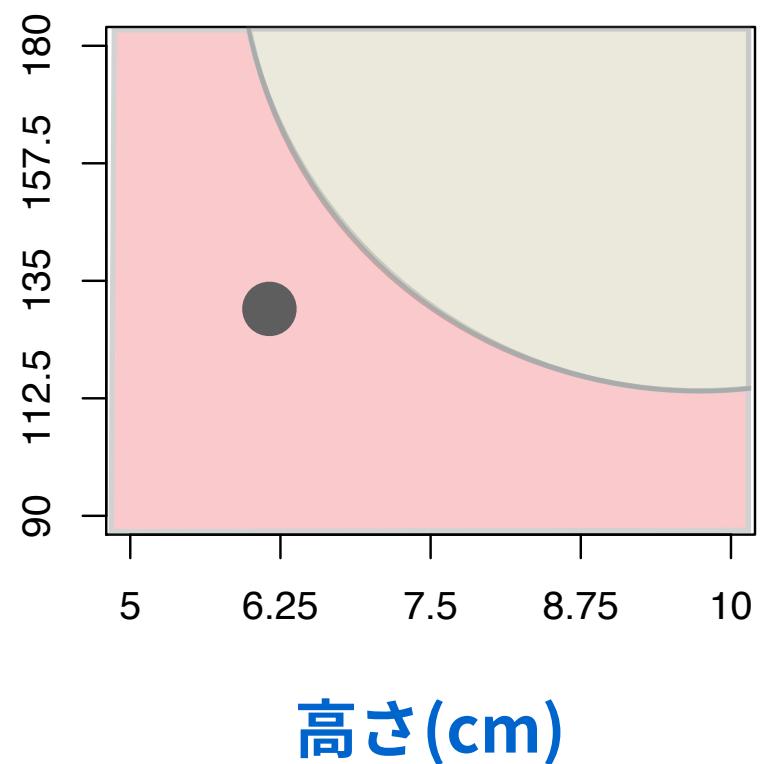
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



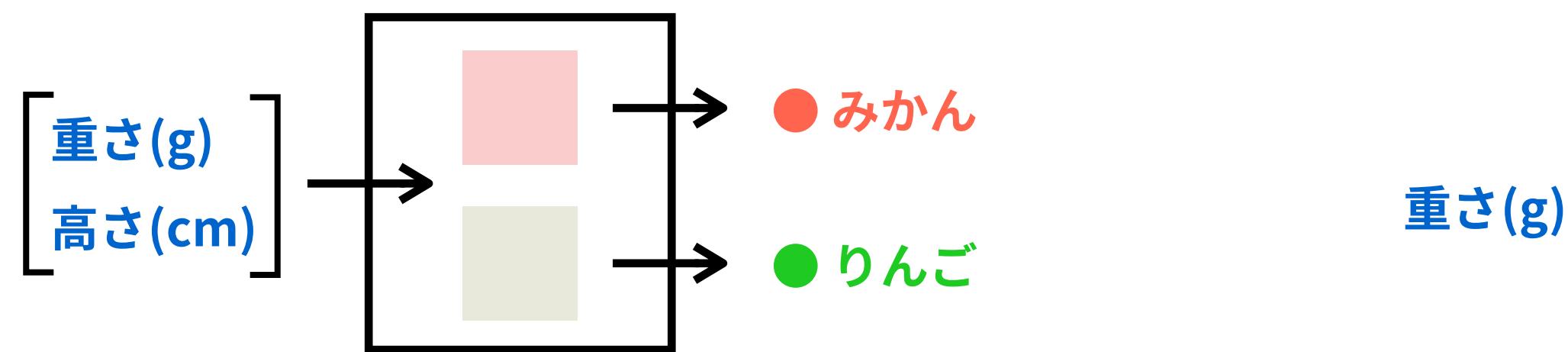
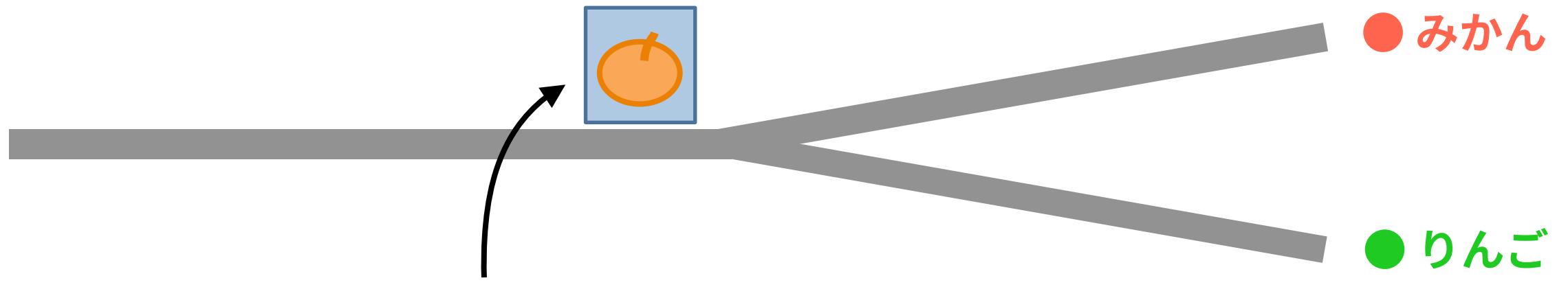
機械学習は「データを予測に変える」



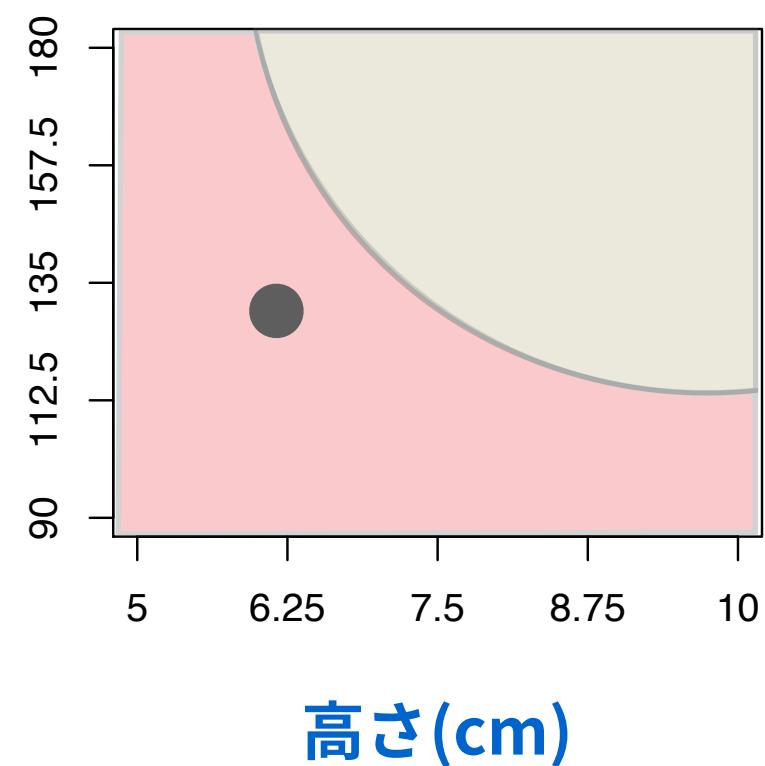
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



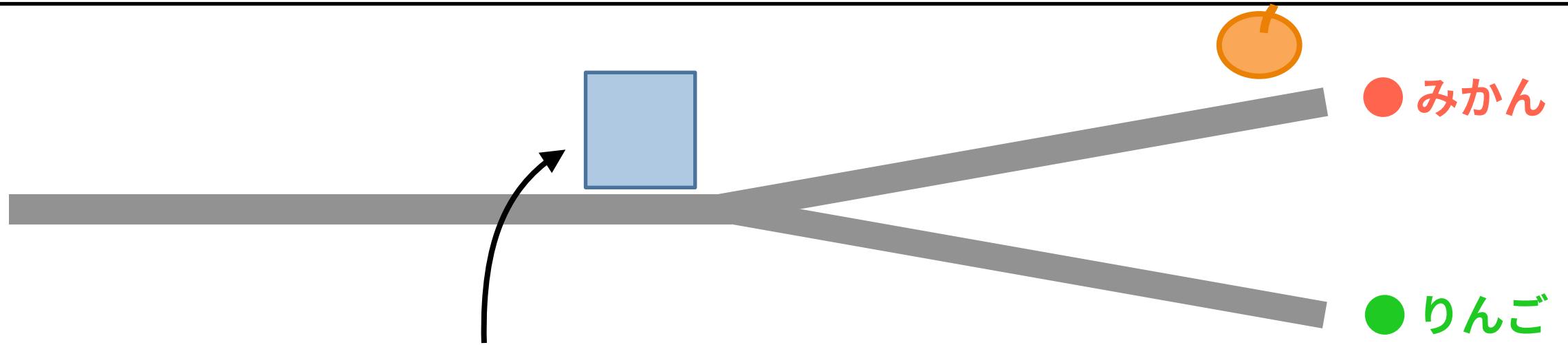
機械学習は「データを予測に変える」



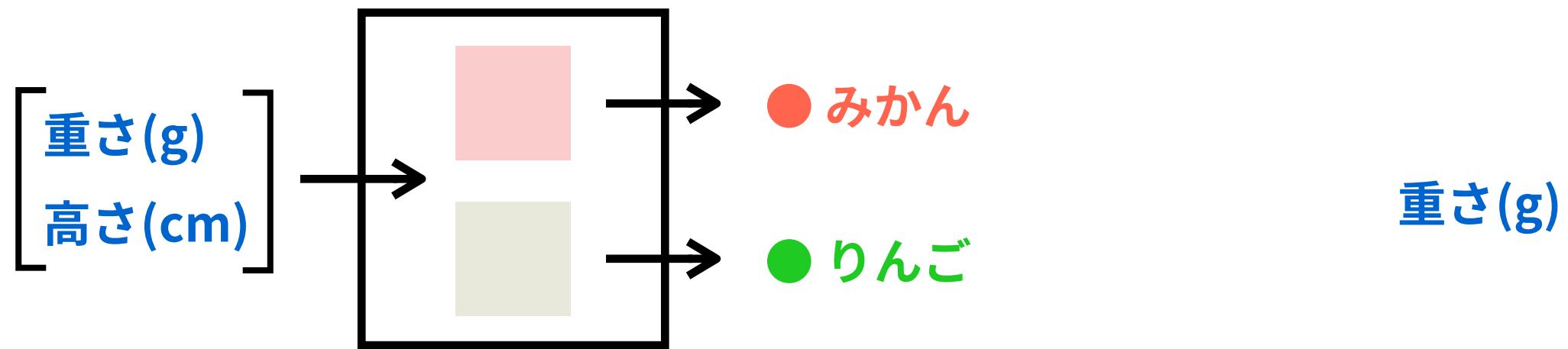
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



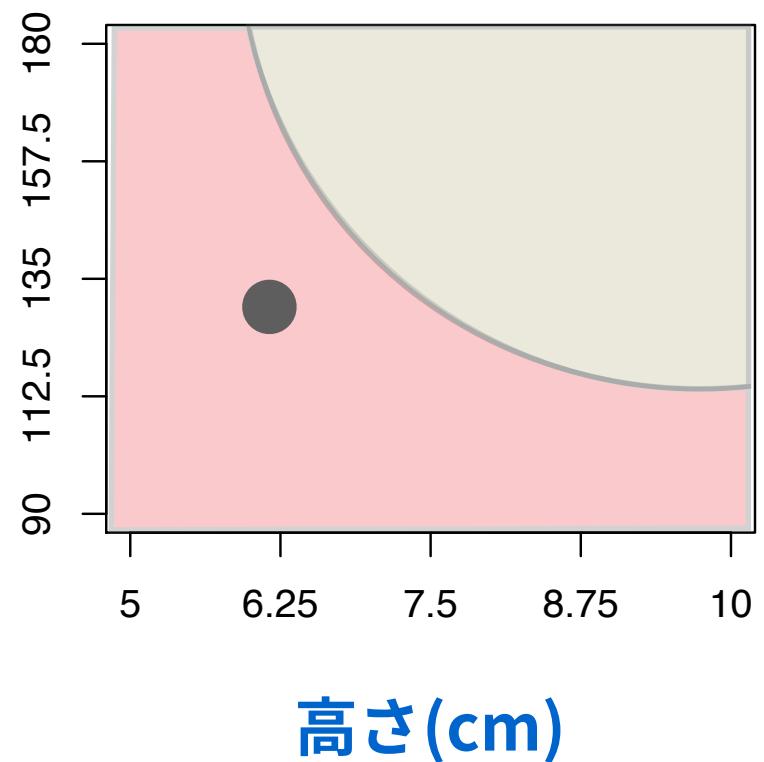
機械学習は「データを予測に変える」



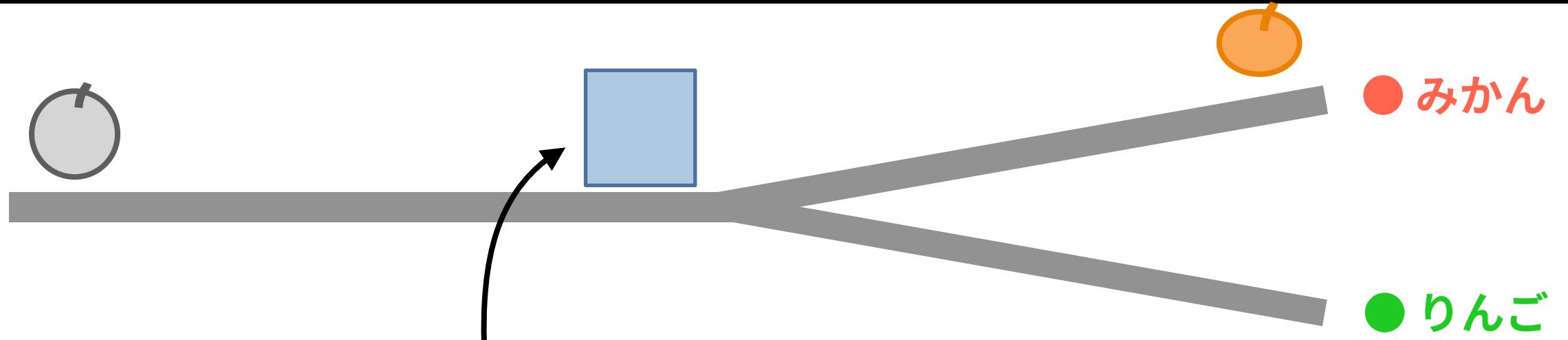
見本データから作っておいた予測プログラム



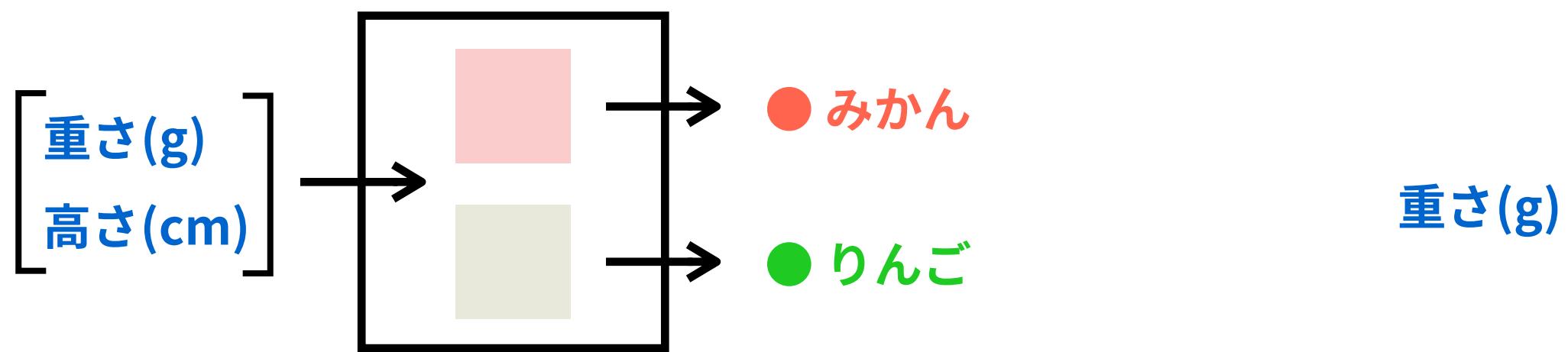
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



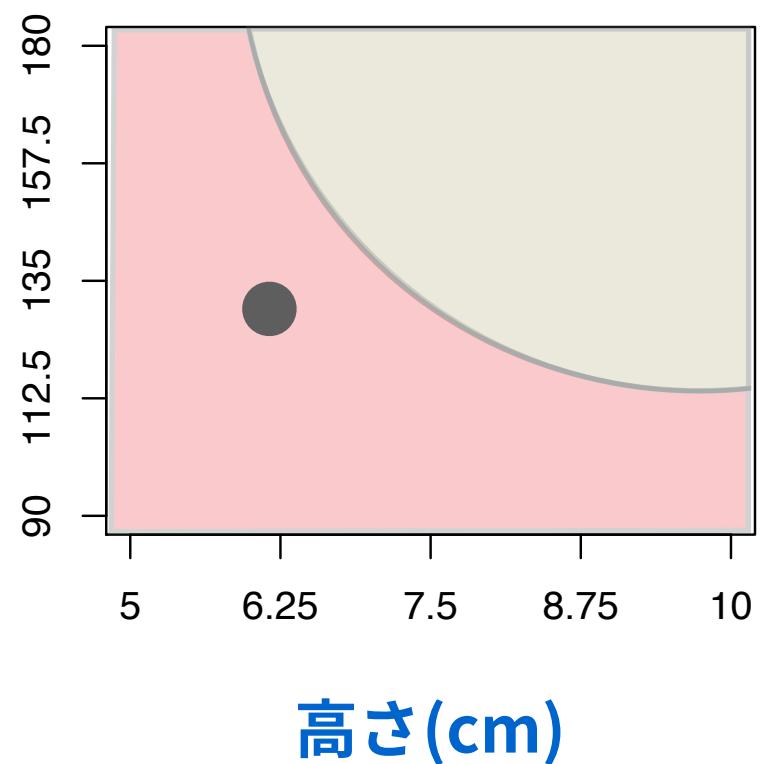
機械学習は「データを予測に変える」



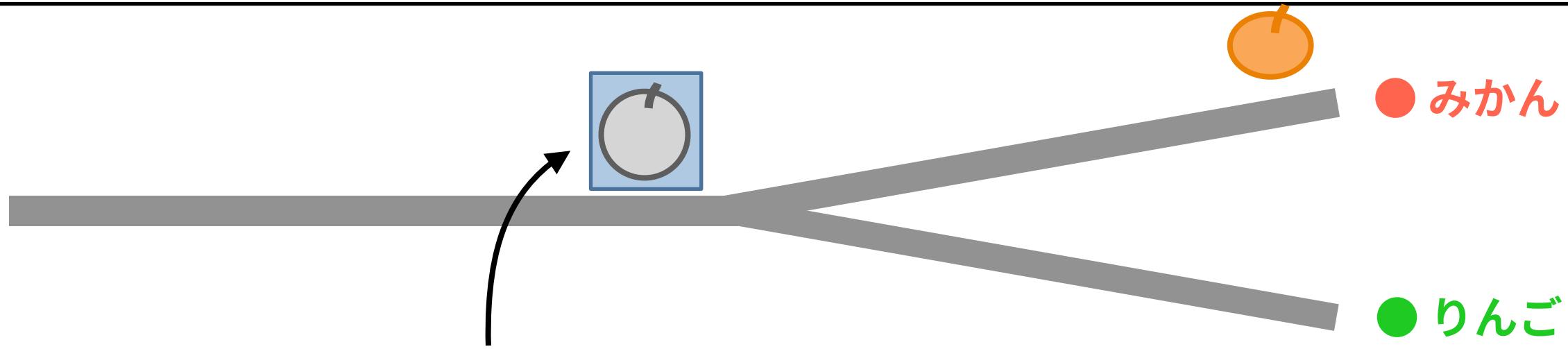
見本データから作っておいた予測プログラム



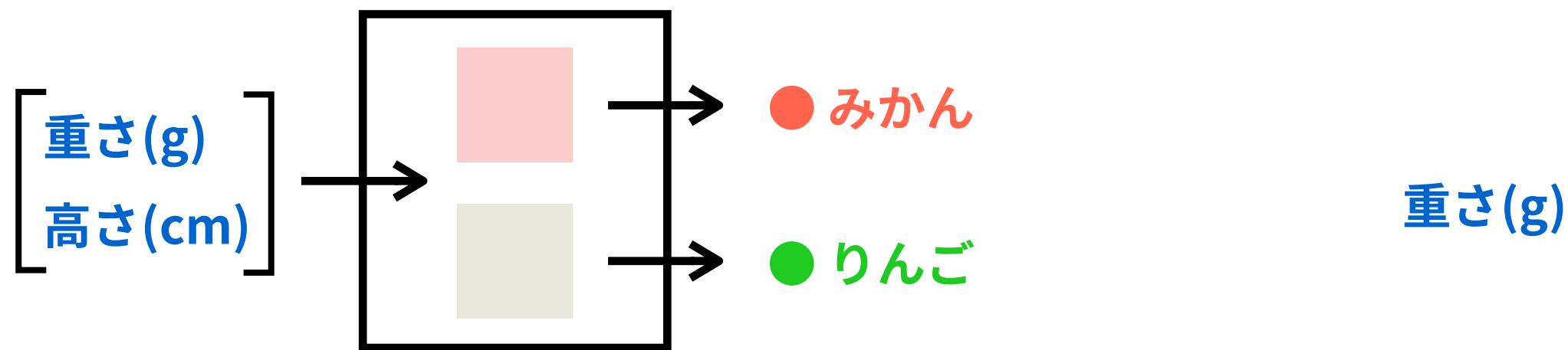
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



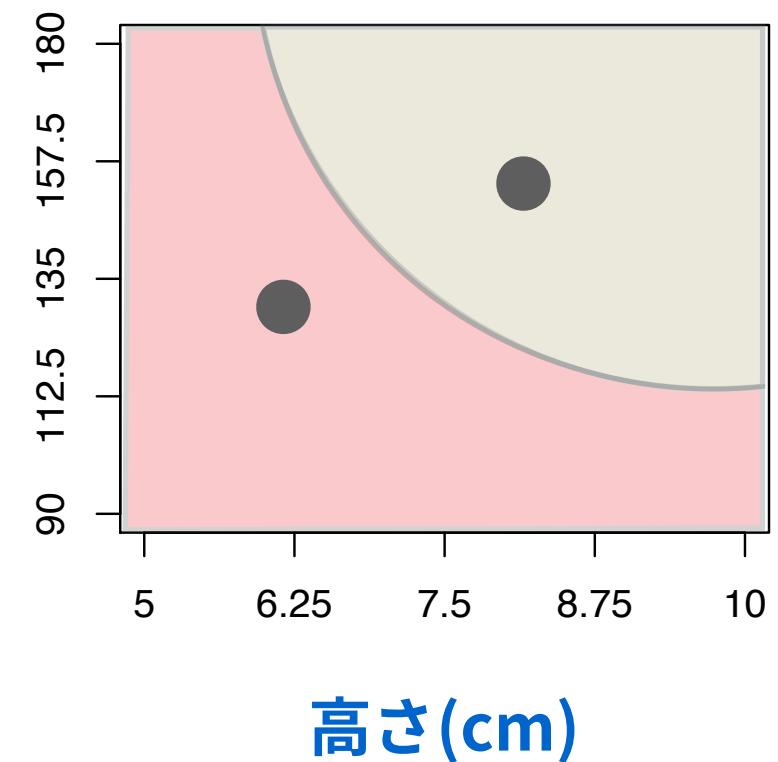
機械学習は「データを予測に変える」



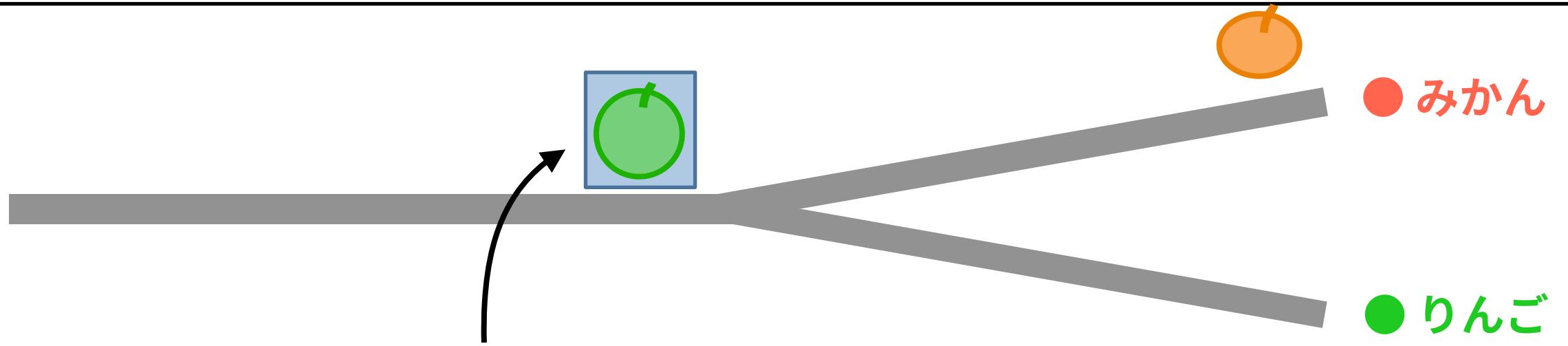
見本データから作っておいた予測プログラム



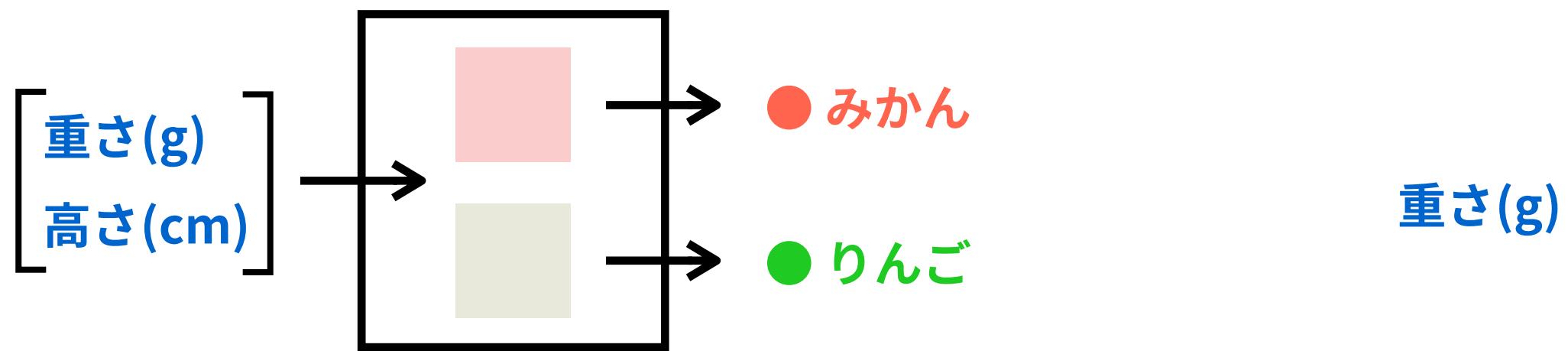
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



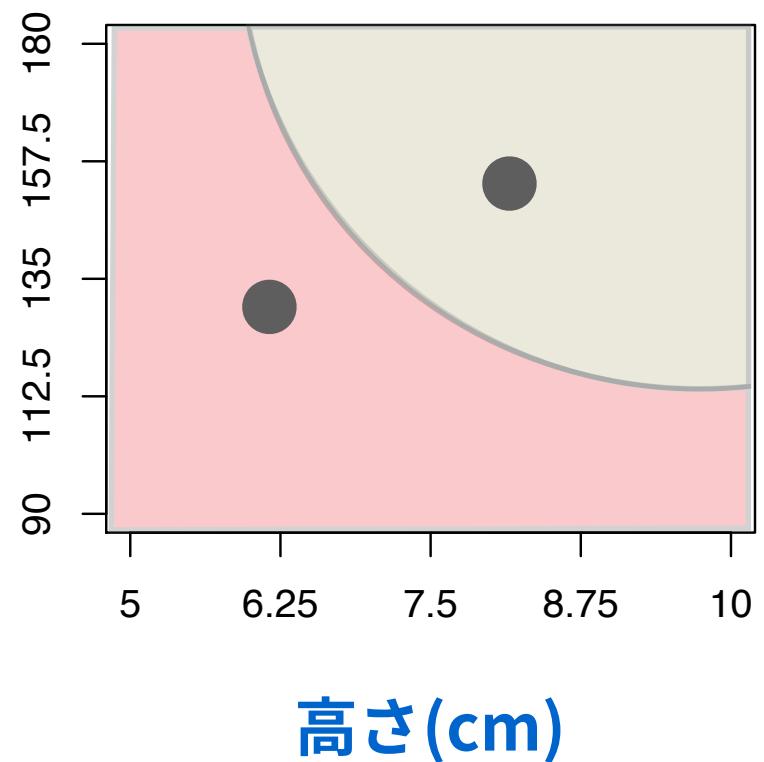
機械学習は「データを予測に変える」



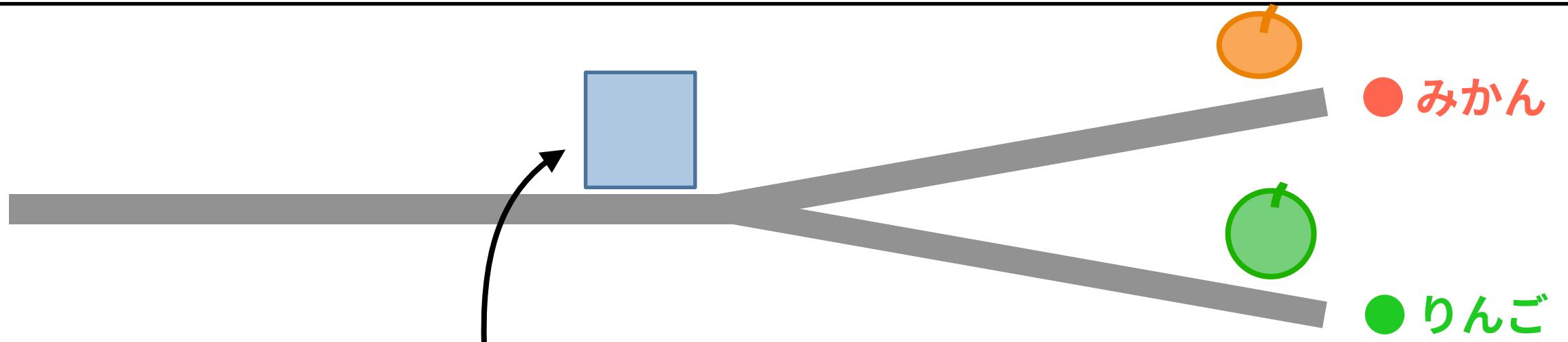
見本データから作っておいた予測プログラム



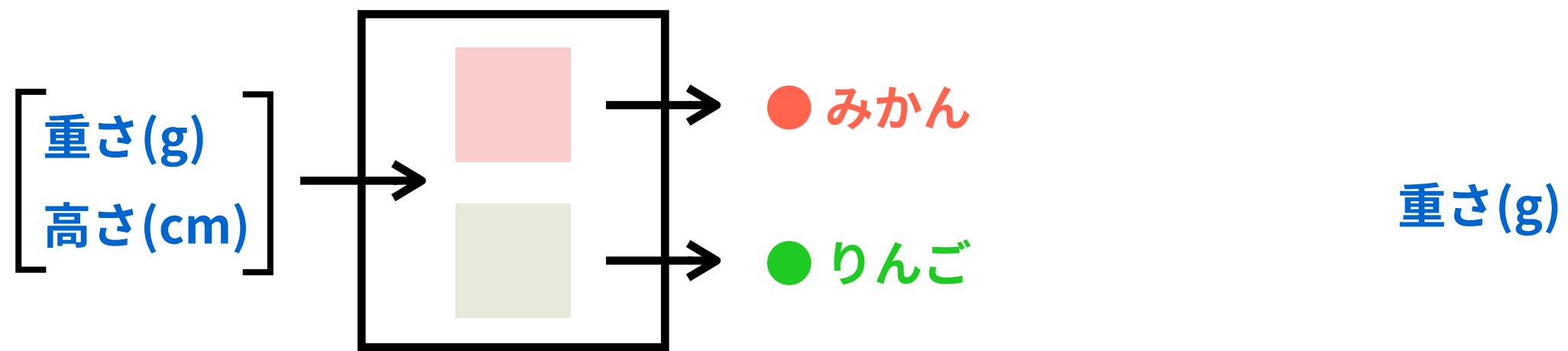
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



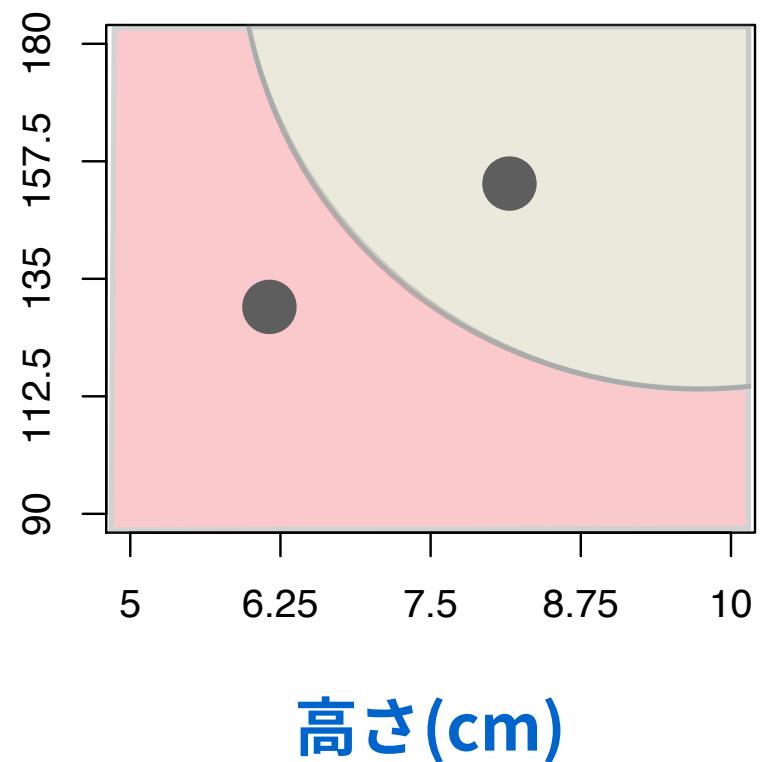
機械学習は「データを予測に変える」



見本データから作っておいた予測プログラム



予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



機械学習は「新しい(雑な)コンピュータプログラムの作り方」

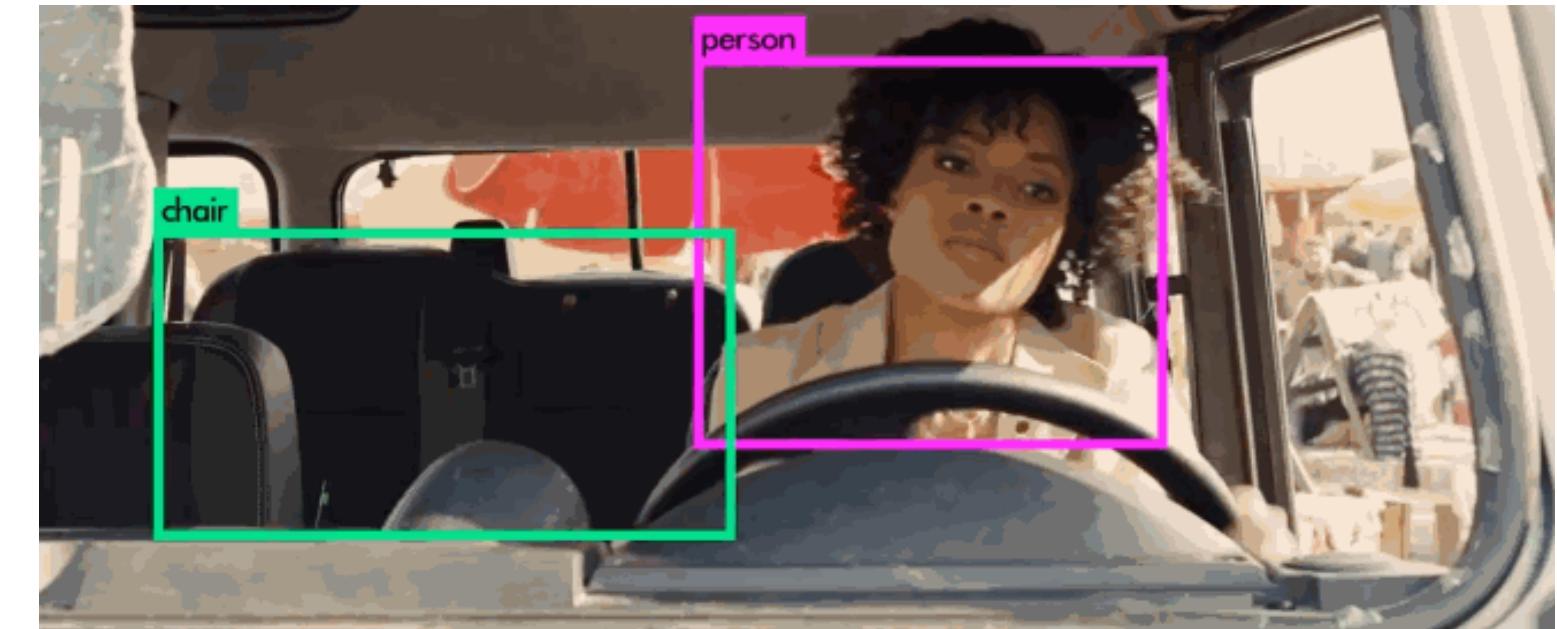
プログラムの入力と出力の関係がよく分からぬ場合でも、

たくさんの入出力の見本データによって間接的にそれを再現できるプログラムを作り出す技術



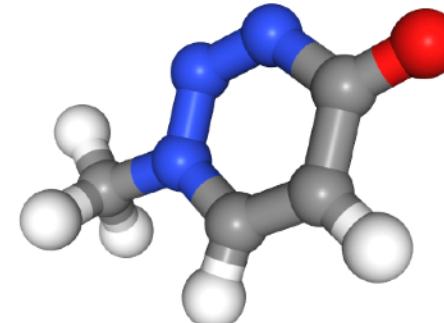
機械学習は「新しい(雑な)コンピュータプログラムの作り方」

この単純なしくみは上手に使うと「めちゃくちゃ強力」でいろいろな楽しいこともできる！



例：機械学習×量子化学

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

量子化学計算

~ 1000 秒

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)

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

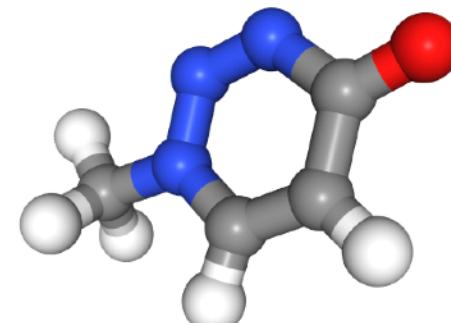
output

	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

- 内部エネルギー
- 自由エネルギー
- ゼロ点振動エネルギー
- 最高被占軌道 (HOMO)
- 最低空軌道 (LUMO)
- 分極率
- 双極子モーメント
- 熱容量
- エンタルピー
- :

例：機械学習×量子化学

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 倍高速！

~ 0.01 秒

機械学習

↔

~ 1000 秒

量子化学計算

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

- 内部エネルギー
- 自由エネルギー
- ゼロ点振動エネルギー
- 最高被占軌道 (HOMO)
- 最低空軌道 (LUMO)
- 分極率
- 双極子モーメント
- 熱容量
- エンタルピー
- :

例：機械學習×量子化学

 MOLSSI Machine Learning Datasets Repository

Search: DFT

Add your Dataset License

Name	↑↓ Quality	↑↓ Data Points	Elements	Sampling	Download
ANI-1	DFT	22,057,374	C H N O	NMS	Download HDF5 Download TEXT
ANI-1x	DFT	4,956,005	C H N O	MD,NMS,DS,TS	Download HDF5
QM9	DFT	133,885	C H F N O	Minima	Download HDF5 Download TEXT

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. & Raymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *JACS*, 2009, 131, 8732-8733. <https://pubs.acs.org/doi/abs/10.1021/ja902302h>
- Ramakrishnan, R.; Dral, P. O.; Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data*, 2014, 1, 140022. <https://www.nature.com/articles/sdata201422>

https://qcarchive.molssi.org/apps/ml_datasets/

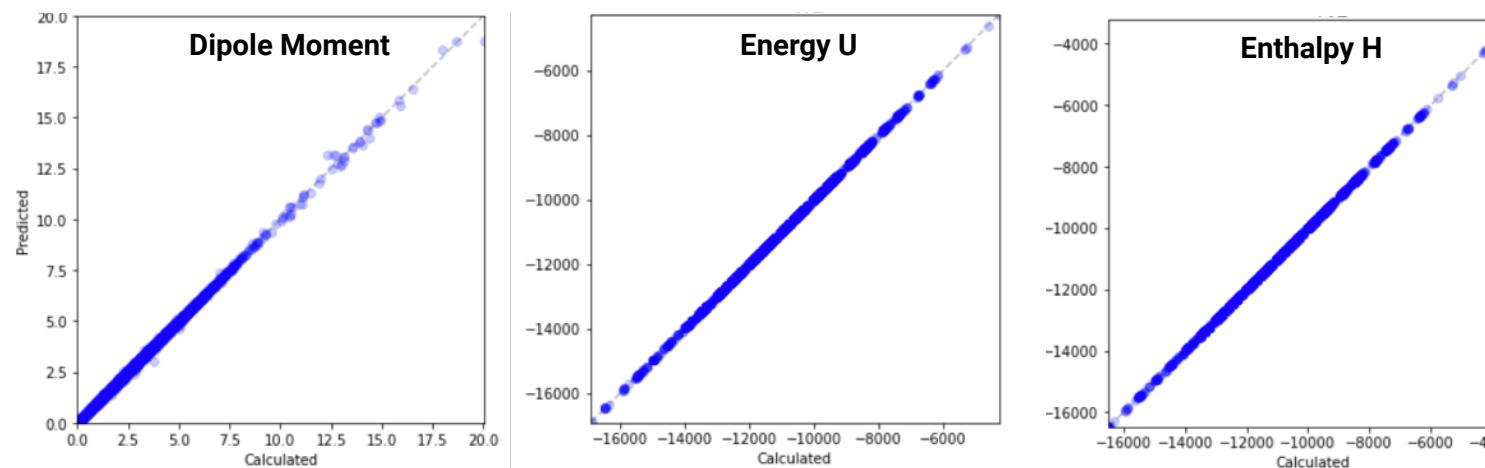
Name	↑↓ Quality	↑↓ Data Points
ANI-1	DFT	22,057,374
ANI-1x	DFT	4,956,005
GDML	DFT, CCSD, CCSD(T)	3,875,468
Solvated Protein Fragments	DFT	2,731,180
ISO-17	DFT	640,982
ANI-1ccx	CCSD(T)*	489,571
SN2 Reactions	DFT	452,709
A Benchmark Data Set for Hydrogen Combustion	wB97X-V/cc-pVTZ	361,803
TensorMol Water Clusters	DFT	354,145
QM9	DFT	133,885
PC9	DFT	99,234
PC9 (neutral singlet subset)	DFT	93,883

Showing 1 to 12 of 23 entries

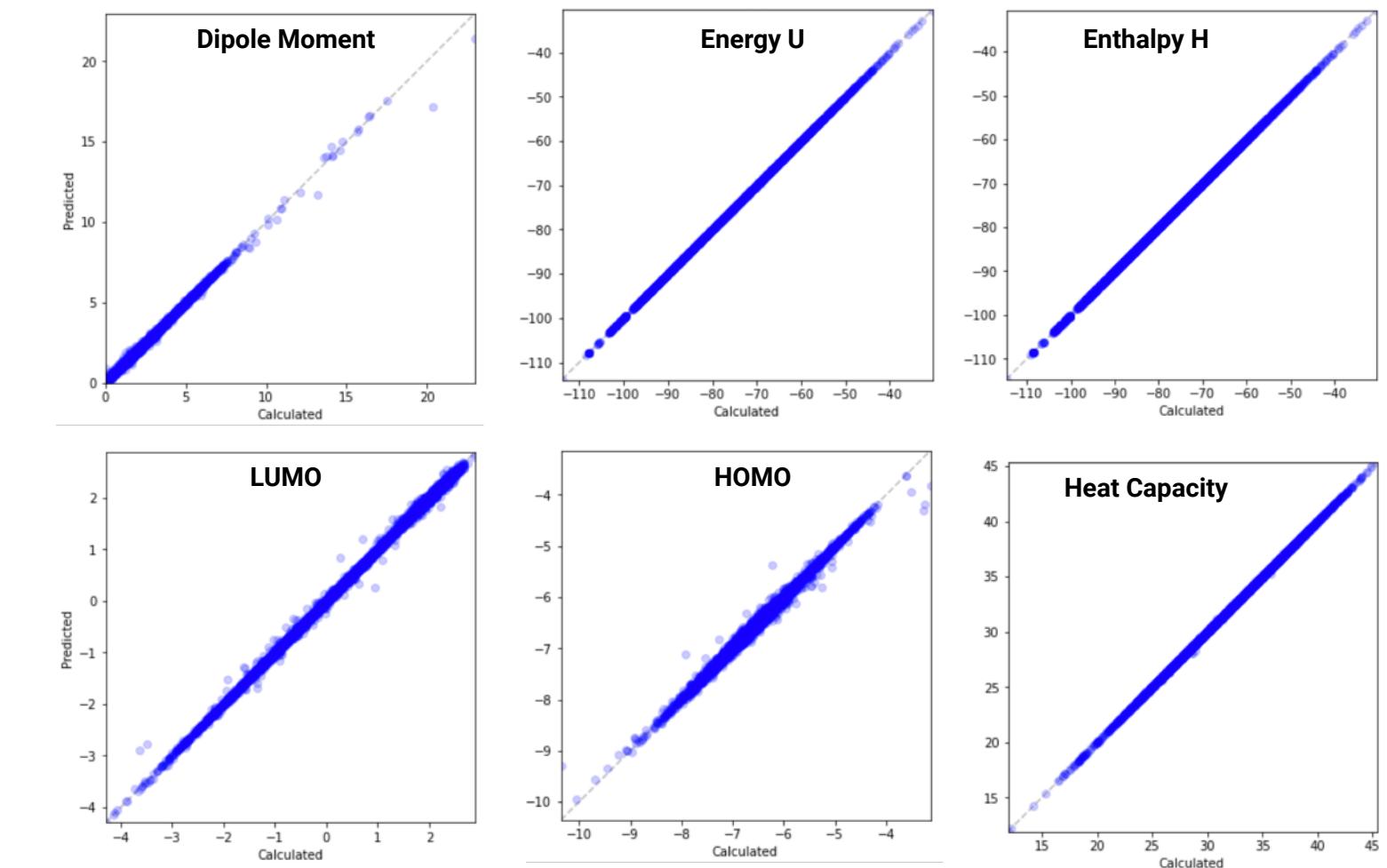
例：機械学習×量子化学

機械学習による予測はめっちゃ当たる！データの揃え方次第では大きな可能性がある

真値(x軸) vs 予測値(y軸) by SchNet (Schütt et al, 2017)



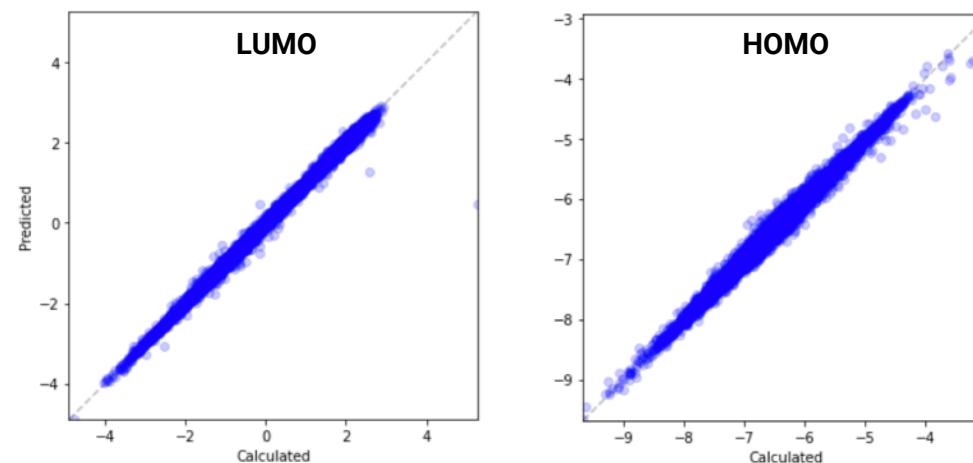
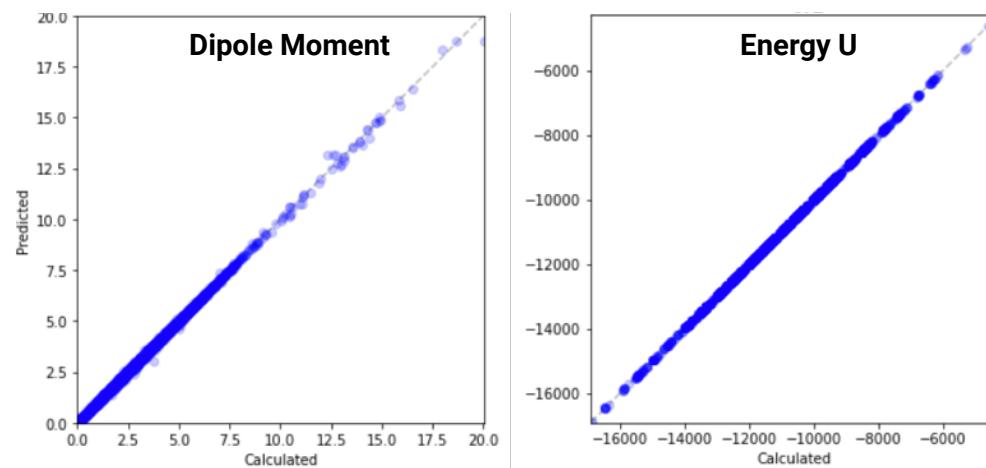
真値(x軸) vs 予測値(y軸) by DimeNet (Klicpera et al, 2020)



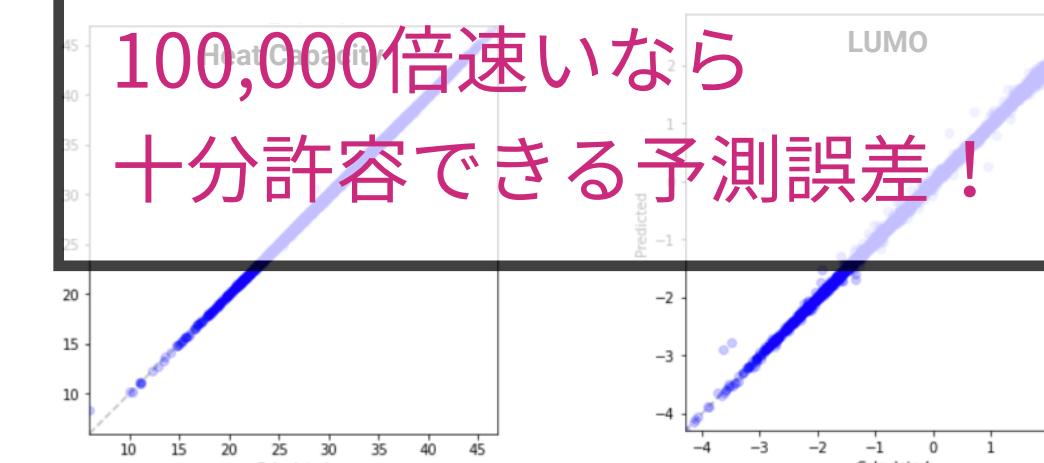
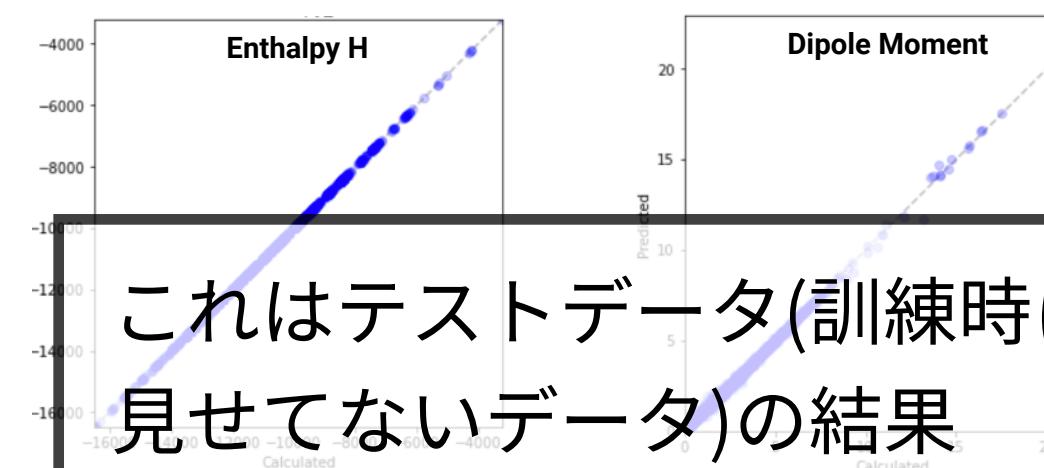
例：機械学習×量子化学

機械学習による予測はめっちゃ当たる！データの揃え方次第では大きな可能性がある

真値(x軸) vs 予測値(y軸) by SchNet (Schütt et al, 2017)

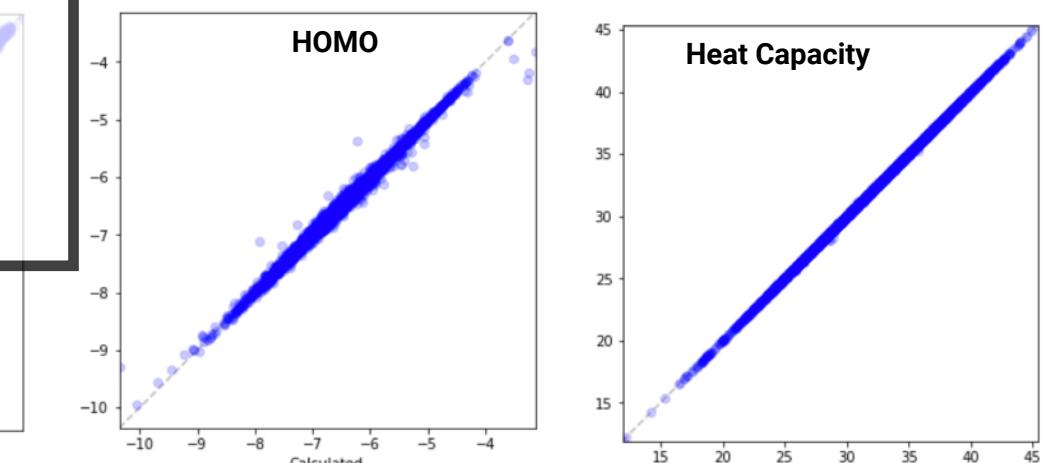
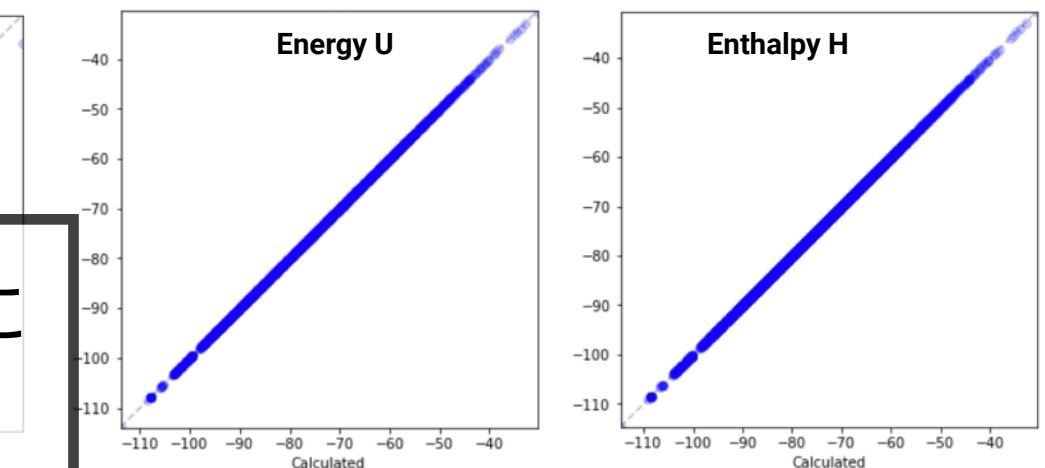


真値(x軸) vs 予測値(y軸) by DimeNet (Klicpera et al, 2020)



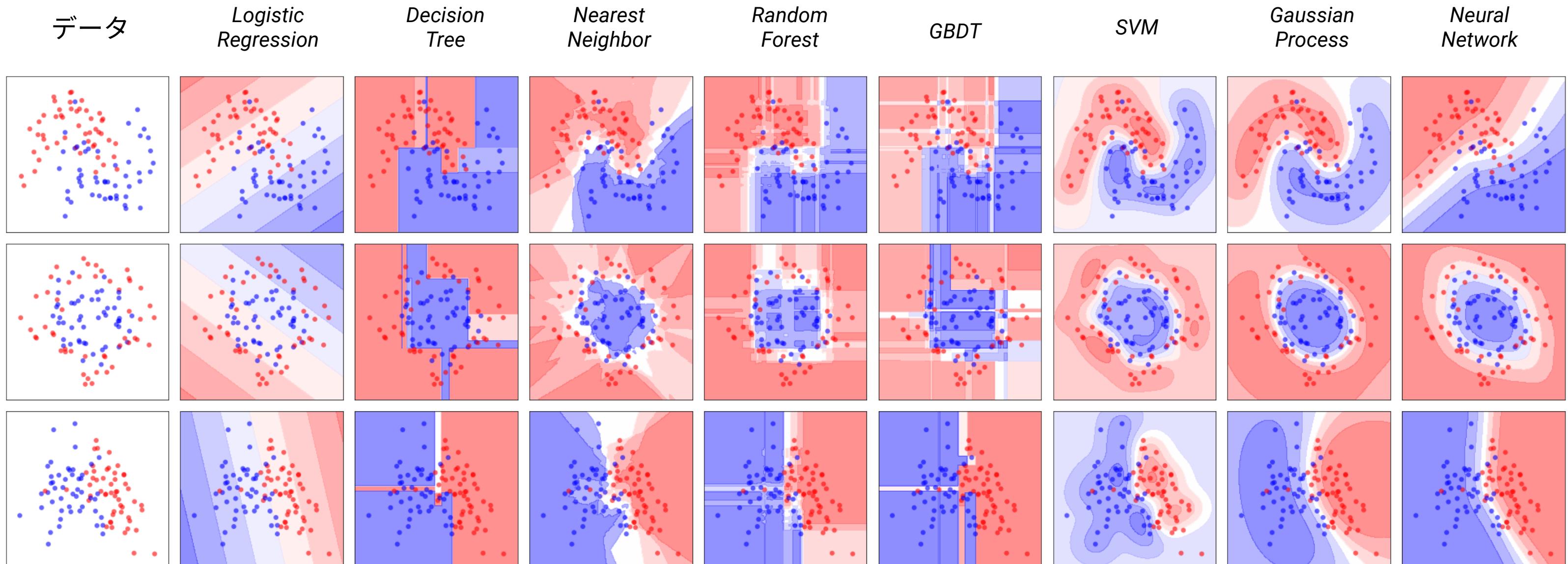
これはテストデータ(訓練時に
見せてないデータ)の結果

100,000倍速いなら
十分許容できる予測誤差！



機械学習は「新しい(雑な)コンピュータプログラムの作り方」

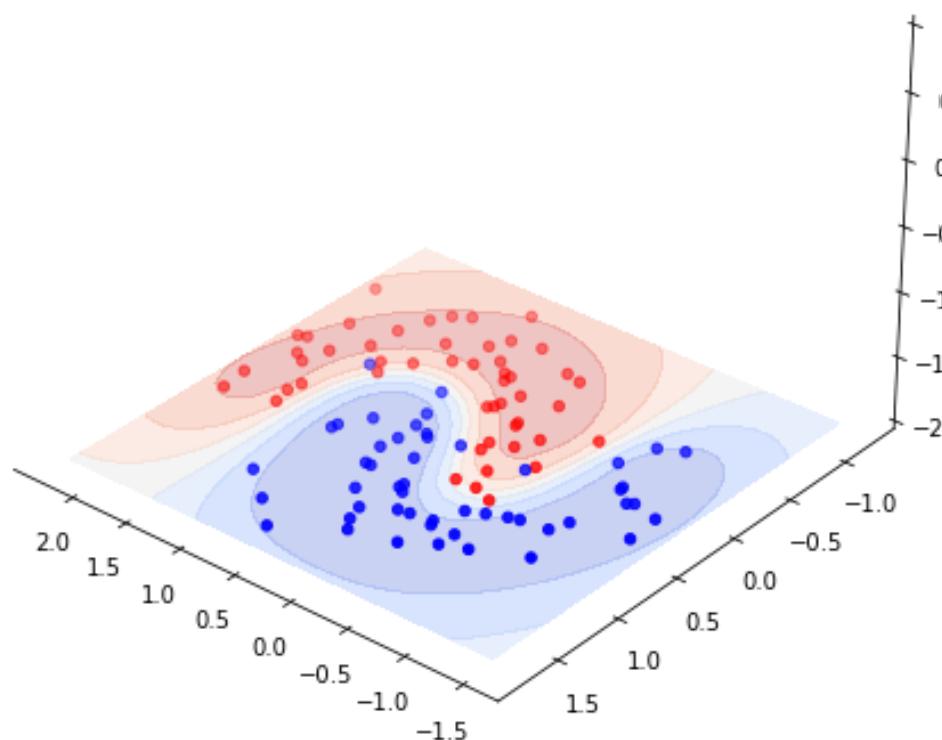
機械学習のアルゴリズムの違いは**境界線の引き方の方針のちがい**だけ



機械学習は「新しい(雑な)コンピュータプログラムの作り方」

内部原理は「曲面モデル」を点にフィッティングしているだけ

→ 境界線の引き方の方針 (帰納バイアス)



機械学習は「新しい(雑な)コンピュータプログラムの作り方」

内部原理は「曲面モデル」を点にフィッティングしているだけ

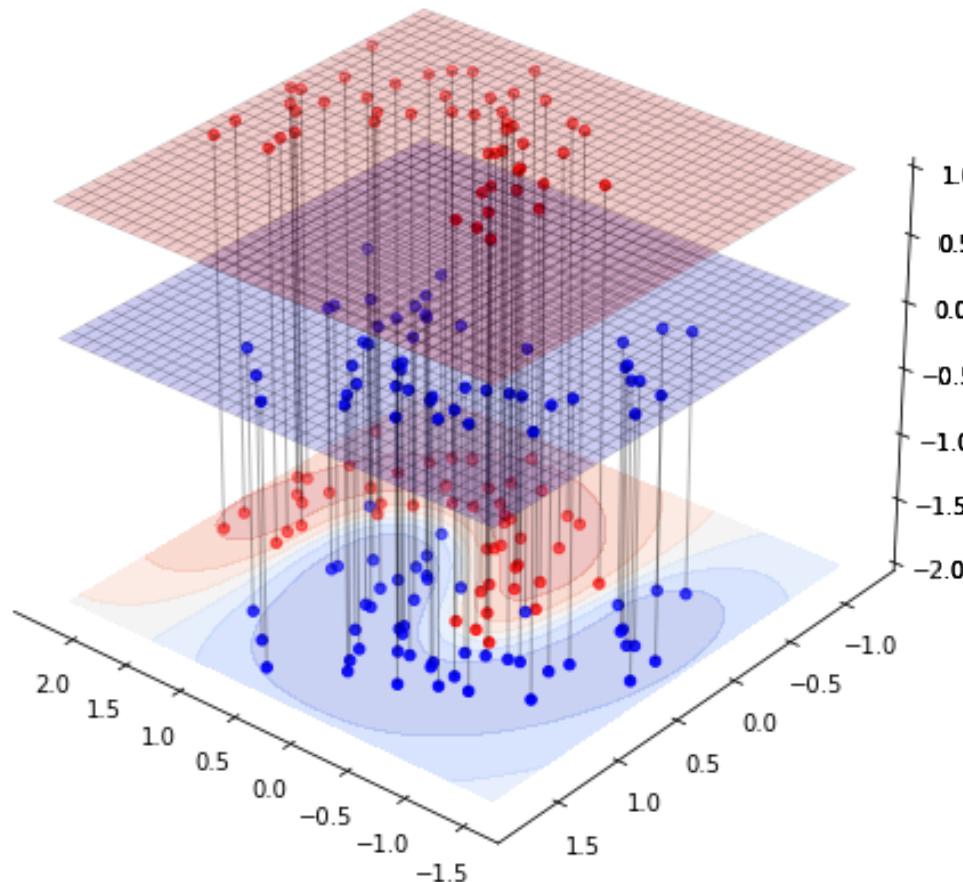
→ 境界線の引き方の方針 (帰納バイアス)



機械学習は「新しい(雑な)コンピュータプログラムの作り方」

内部原理は「曲面モデル」を点にフィッティングしているだけ

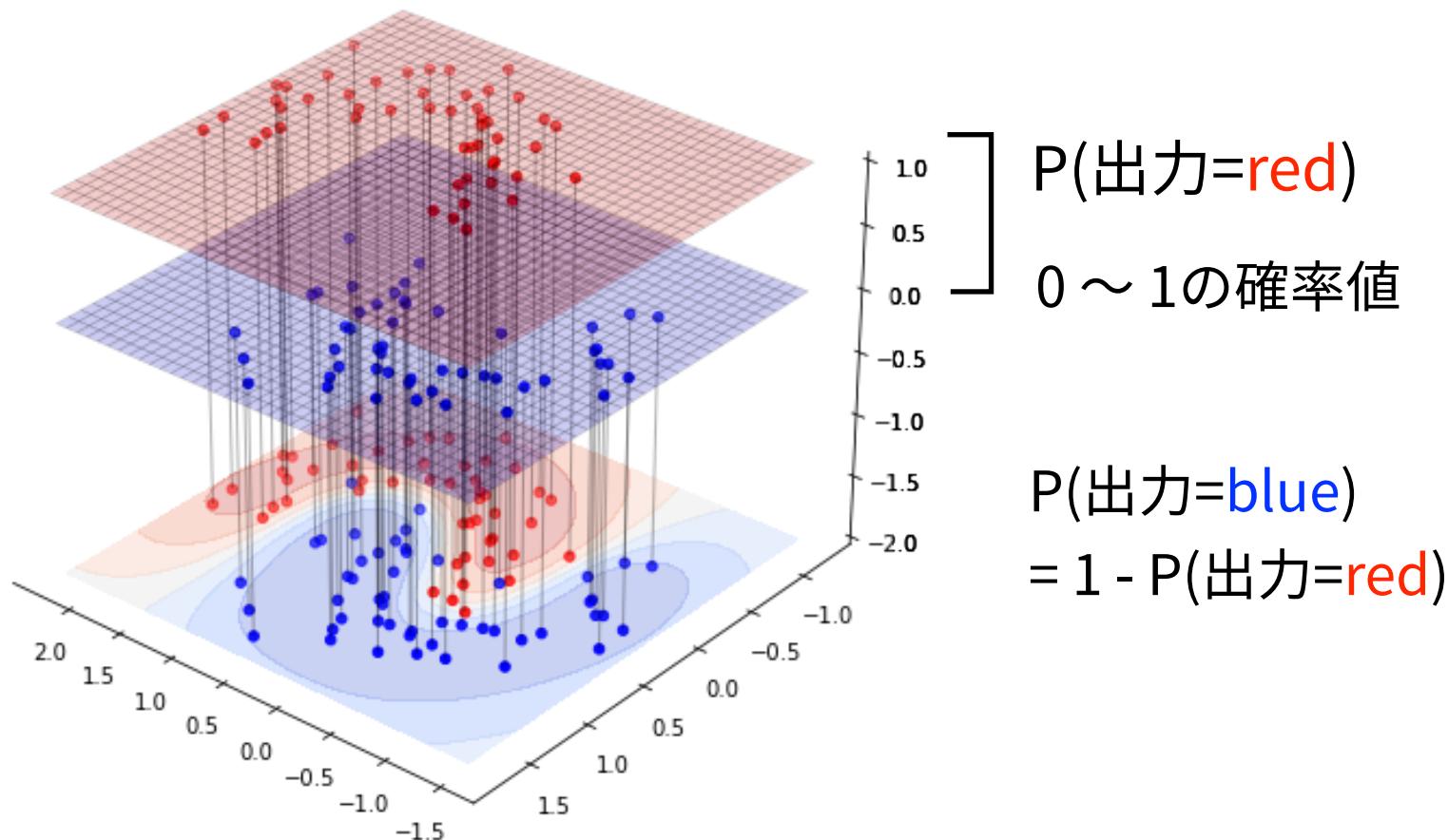
→ 境界線の引き方の方針 (帰納バイアス)



機械学習は「新しい(雑な)コンピュータプログラムの作り方」

内部原理は「曲面モデル」を点にフィッティングしているだけ

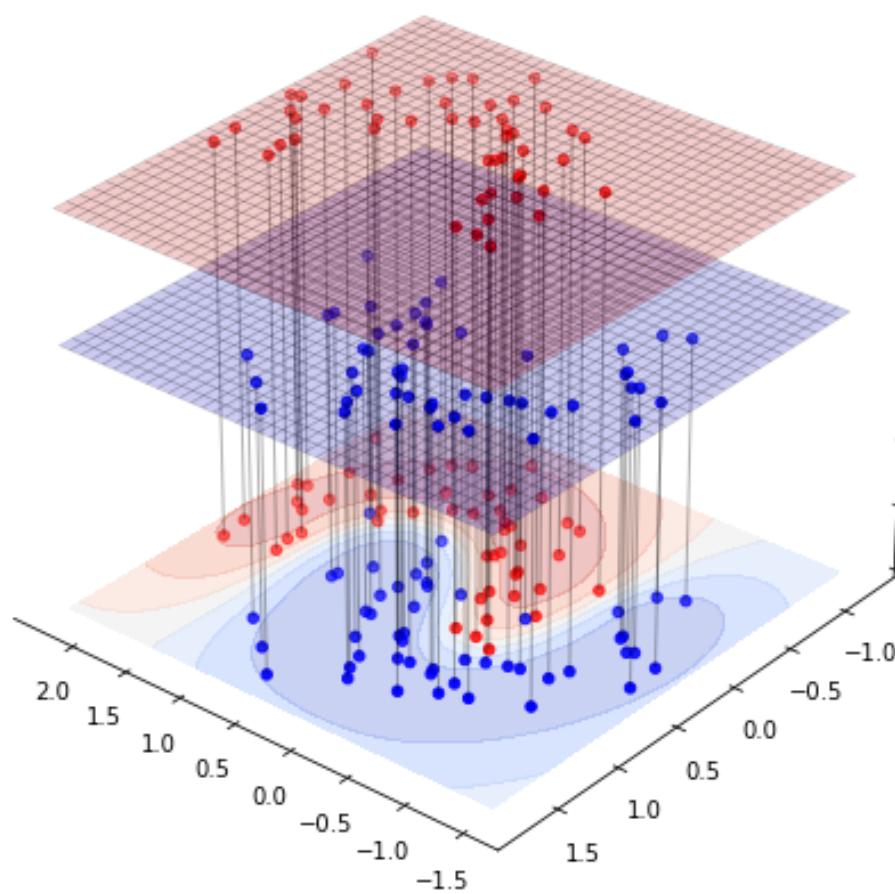
→ 境界線の引き方の方針 (帰納バイアス)



機械学習は「新しい(雑な)コンピュータプログラムの作り方」

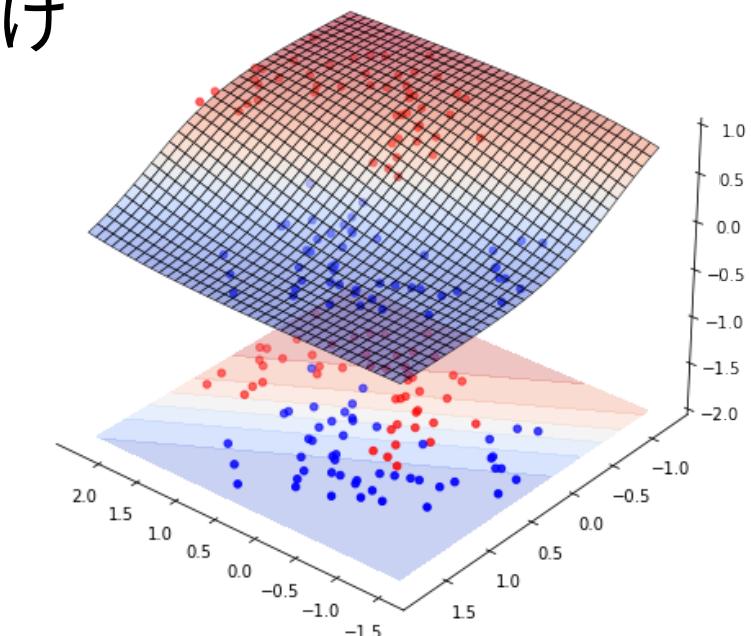
内部原理は「曲面モデル」を点にフィッティングしているだけ

→ 境界線の引き方の方針 (帰納バイアス)

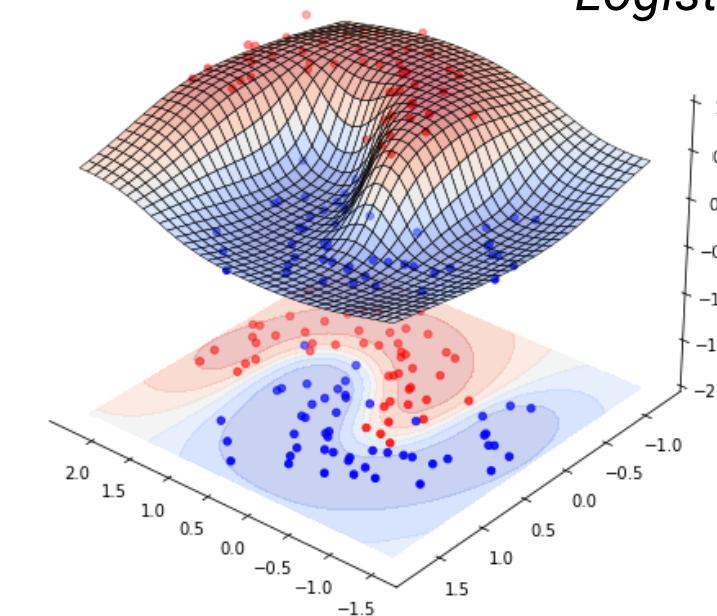


$P(\text{出力}=\text{red})$
0 ~ 1 の確率値

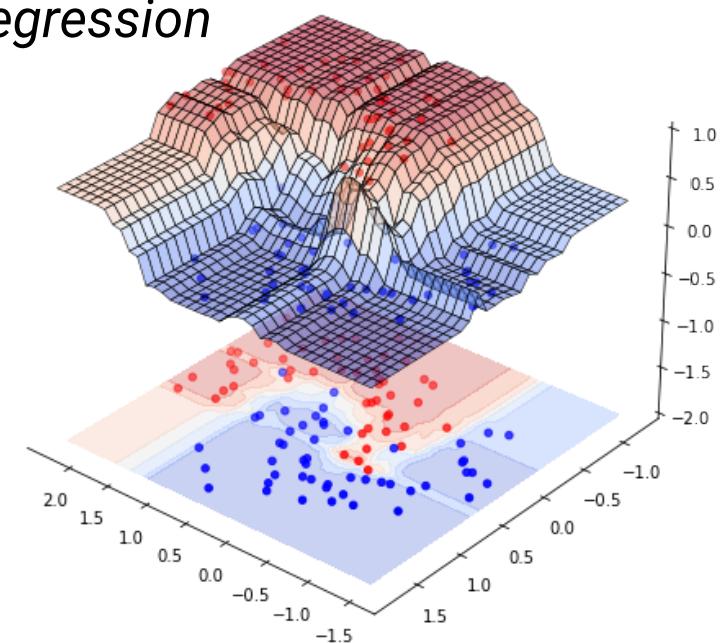
$$P(\text{出力}=\text{blue}) = 1 - P(\text{出力}=\text{red})$$



Logistic Regression



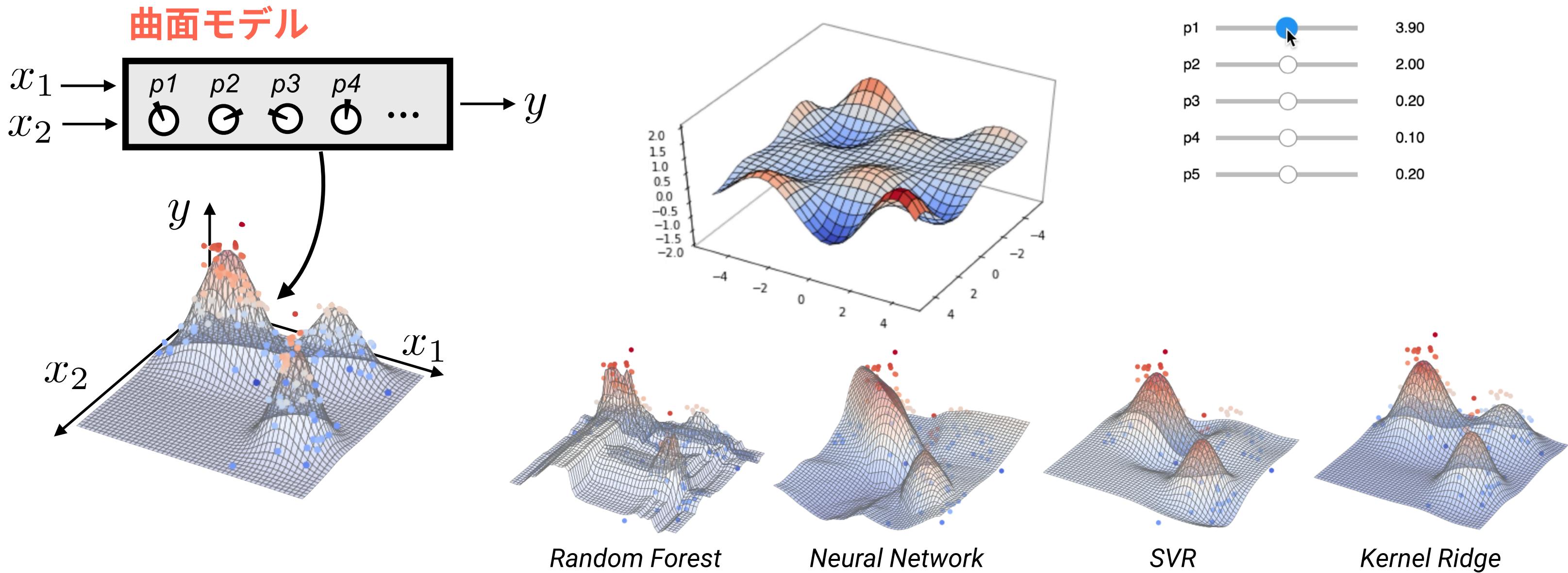
Gaussian Process



Random Forest

機械学習は「新しい(雑な)コンピュータプログラムの作り方」

「曲面モデル」の内部パラメタ値を調整して見本点にあうようフィッティングする



蛇足：希望的呼び方による幻想にご注意を

現在の機械学習は一般に連想されるSF的な「人工知能(AI)」とはかなりかけ離れているが、
「データを予測に変える」側面があまりに強力なため、私たちの日常生活から今後の社会の
カタチにまで影響を及ぼそうとしている…

wishful mnemonics

「人工知能」「機械学習」などの希望的呼び方は本質をミスリードしやすいのでご注意を！

https://spectrum.ieee.org/files/11920/10_Spectrum_2021.pdf

What's Next for Deep Learning > Another AI winter or eternal sunshine? P. 26

Inside DeepMind's Robot Lab > An AI powerhouse takes on "catastrophic forgetting" P. 34

The 7 Biggest Weaknesses of Neural Nets > Surprise! One of them is math P. 42

FOR THE TECHNOLOGY INSIDER

OCTOBER 2021

IEEE Spectrum

Why Is AI So Dumb?

A SPECIAL REPORT

SIGART Newsletter No. 57 April 1976

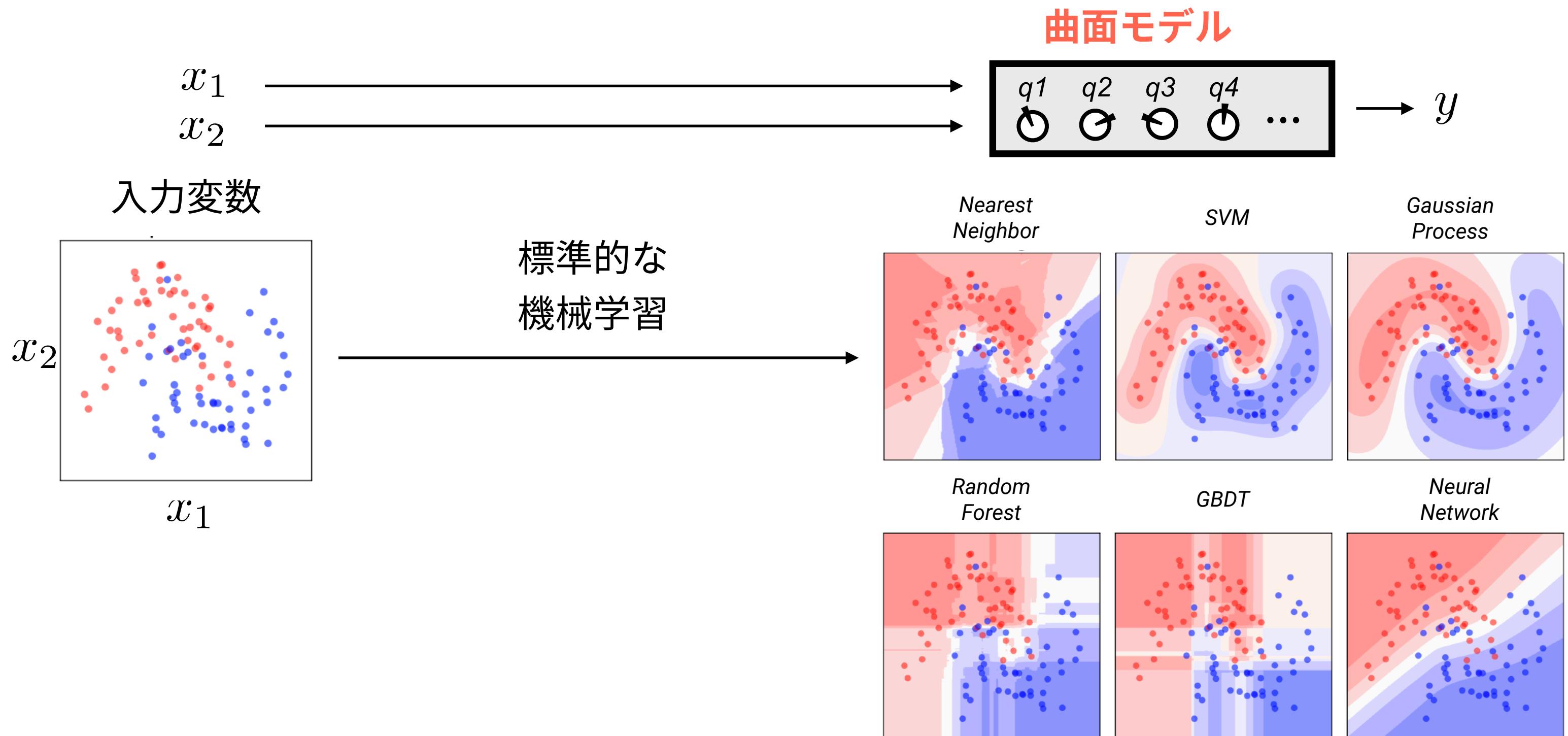
ARTIFICIAL INTELLIGENCE MEETS NATURAL STUPIDITY

Drew McDermott

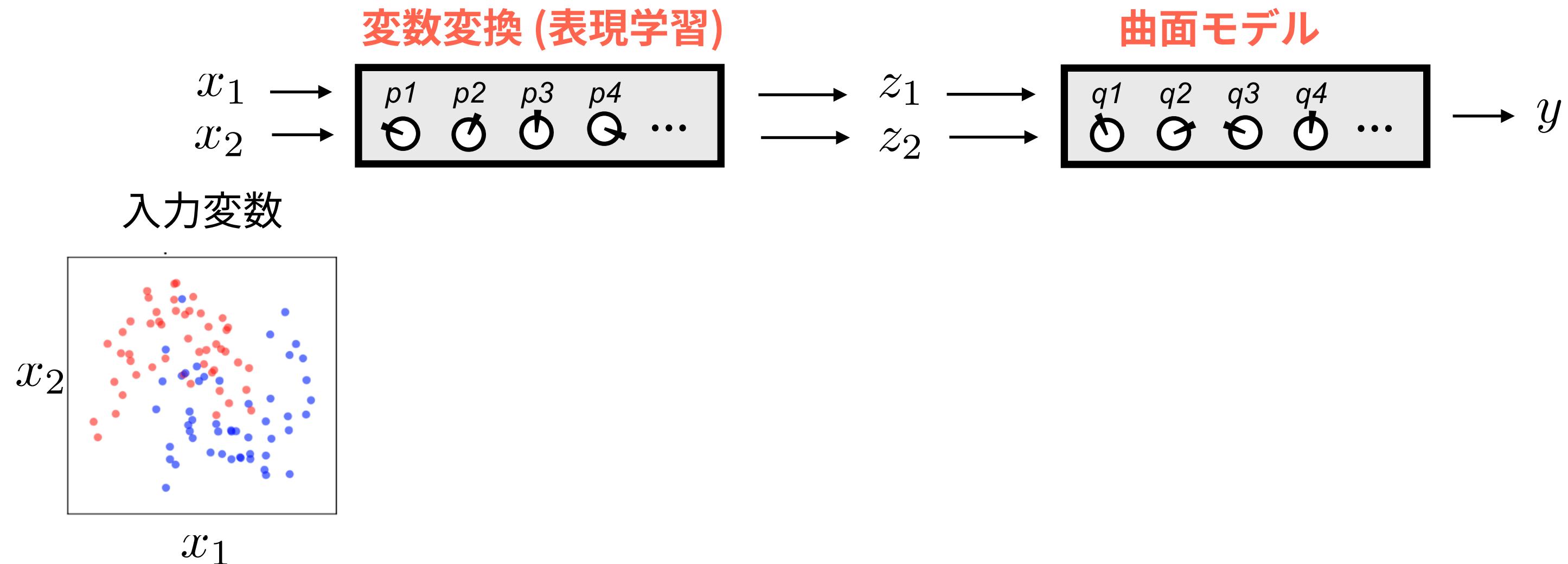
MIT AI Lab Cambridge, Mass 02139

As a field, artificial intelligence has always been on the border of respectability, and therefore on the border of crackpottery. Many critics <Dreyfus, 1972>, <Lighthill, 1973> have urged that we are over the border. We have been very defensive toward this charge, drawing ourselves up with dignity when it is made and folding the cloak of Science about us. On the other hand, in private, we have been justifiably proud of our willingness to explore weird ideas, because pursuing them is the only way to make progress.

深層学習と表現学習

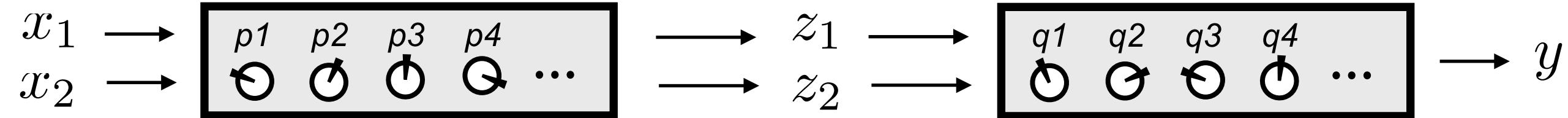


深層学習と表現学習



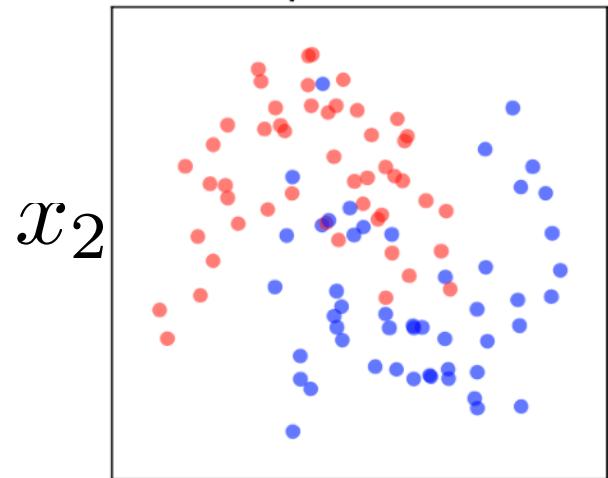
深層学習と表現学習

変数変換(表現学習)

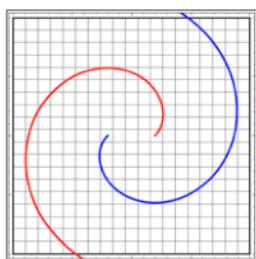


曲面モデル

入力変数

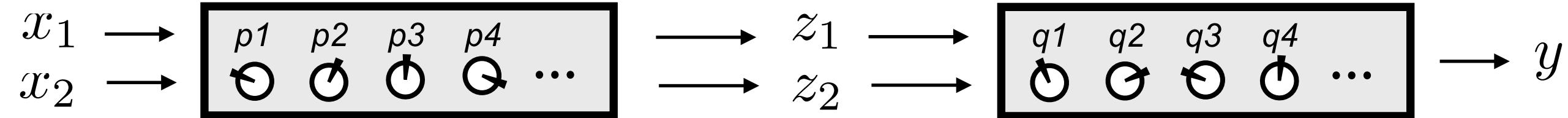


x_1



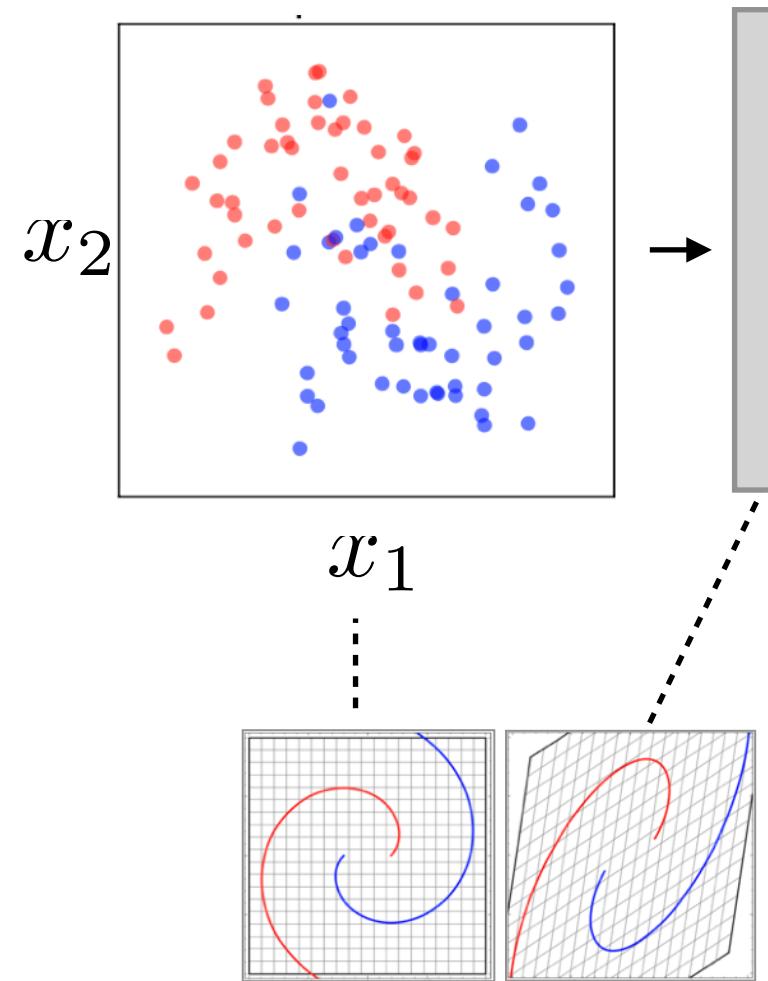
深層学習と表現学習

変数変換 (表現学習)



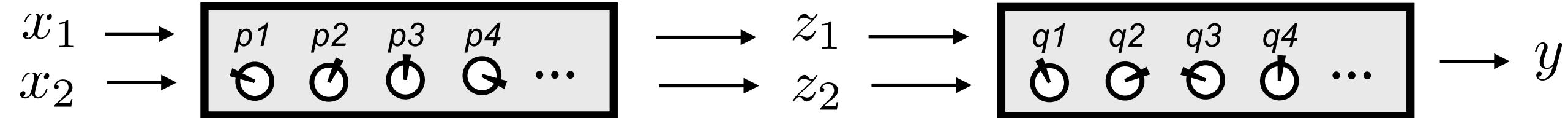
曲面モデル

入力変数

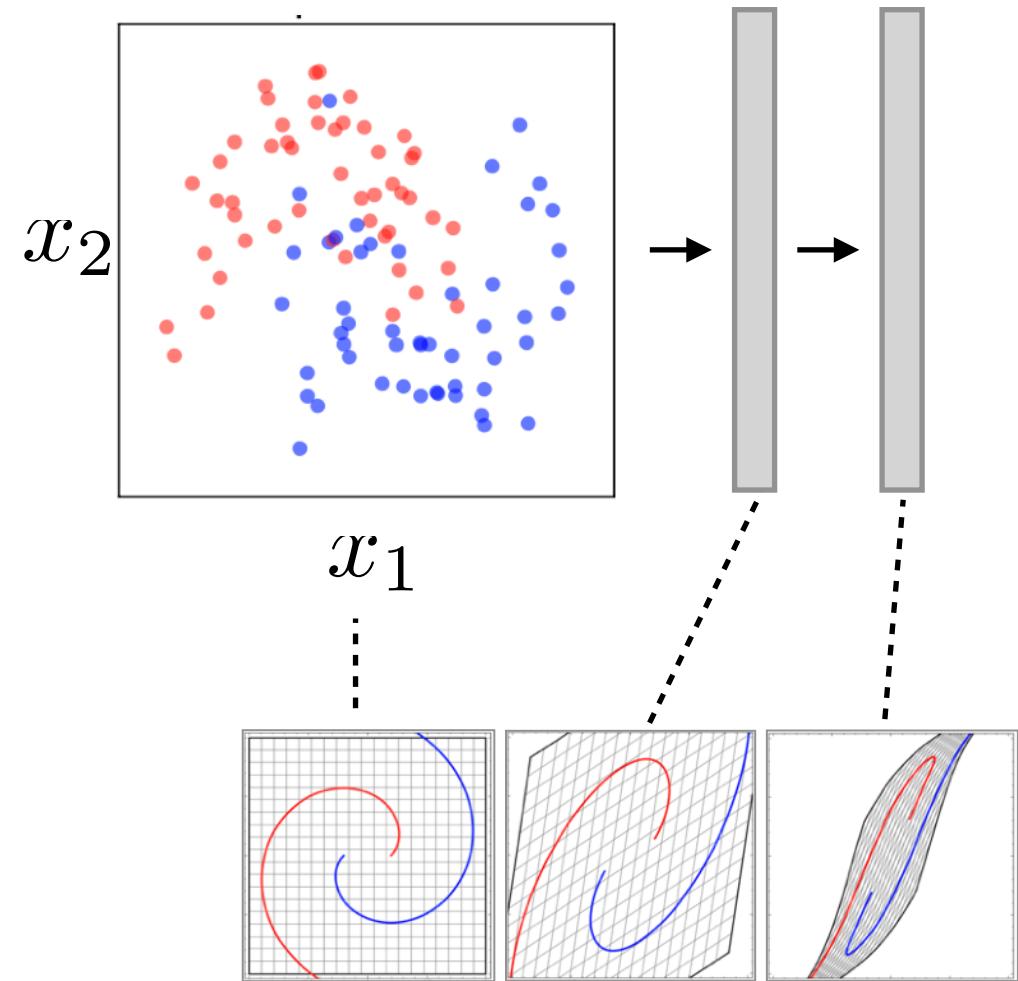


深層学習と表現学習

変数変換 (表現学習)

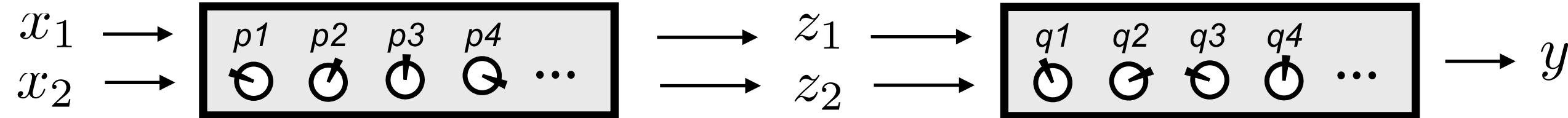


入力変数



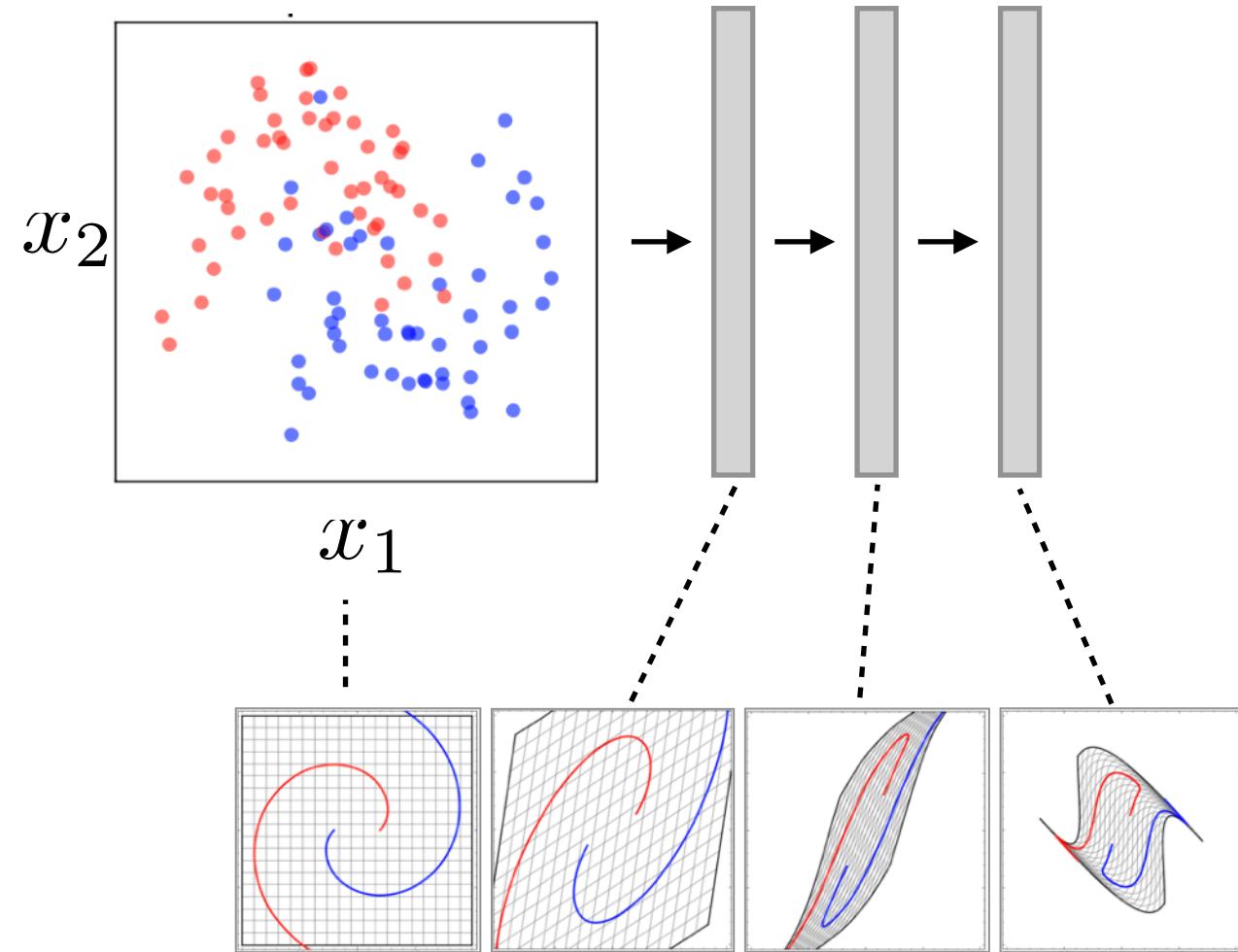
深層学習と表現学習

変数変換 (表現学習)



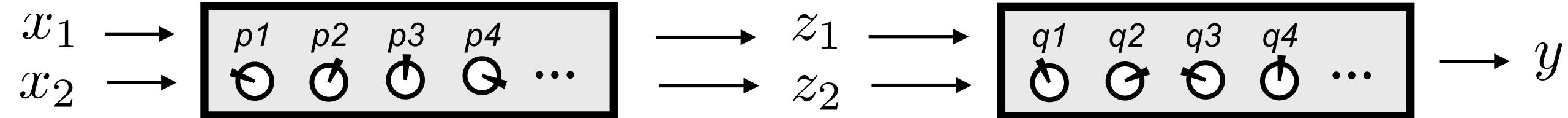
曲面モデル

入力変数



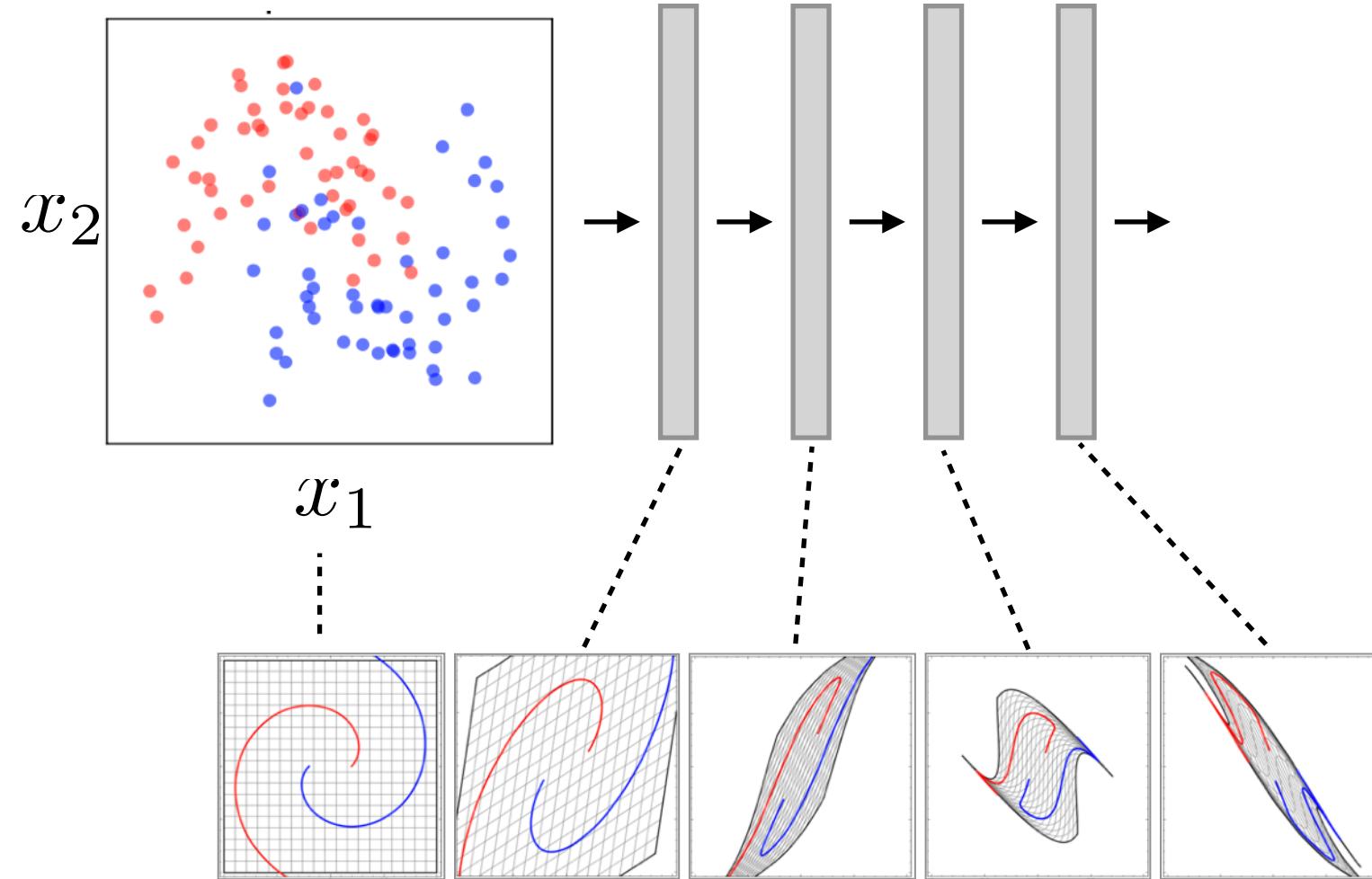
深層学習と表現学習

変数変換 (表現学習)



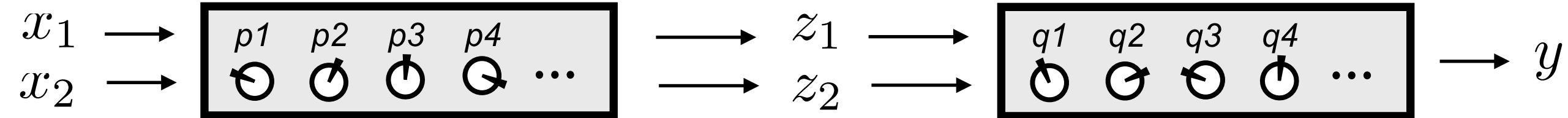
曲面モデル

入力変数

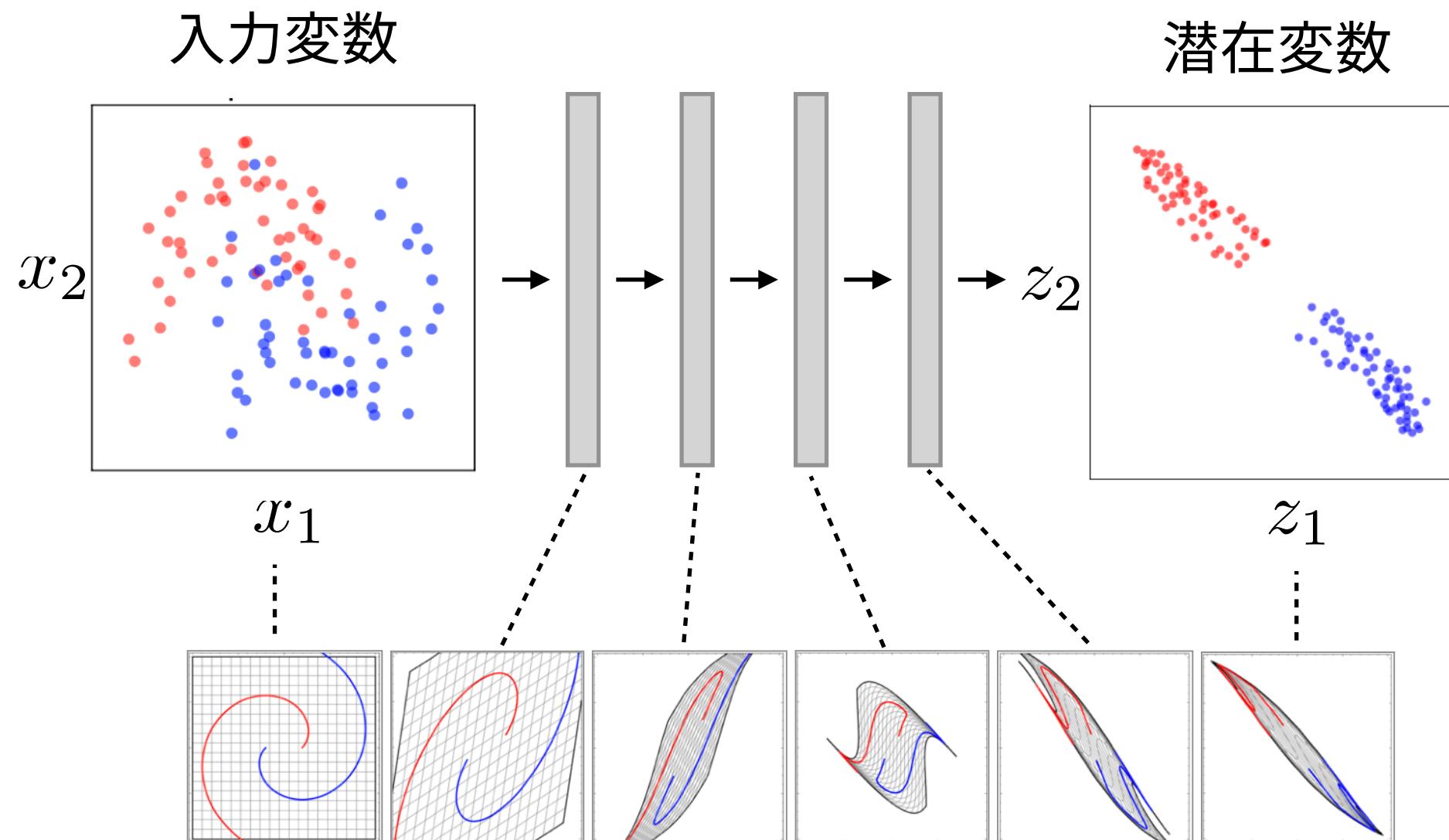


深層学習と表現学習

変数変換(表現学習)

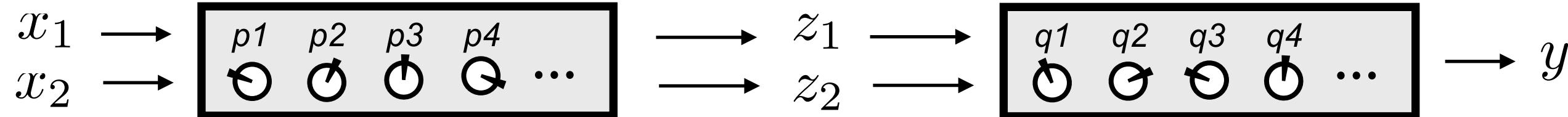


曲面モデル

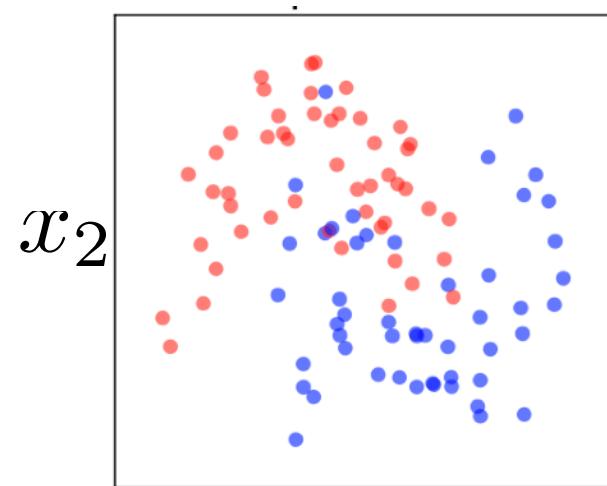


深層学習と表現学習

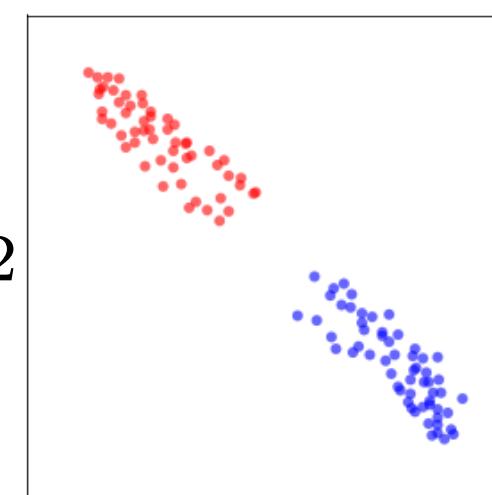
変数変換(表現学習)



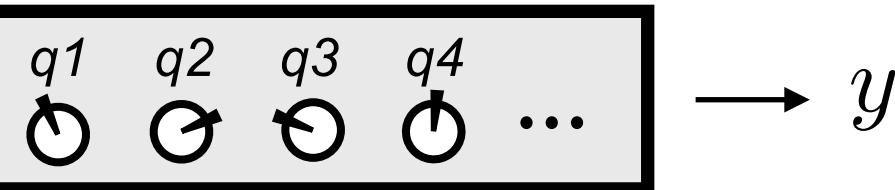
入力変数



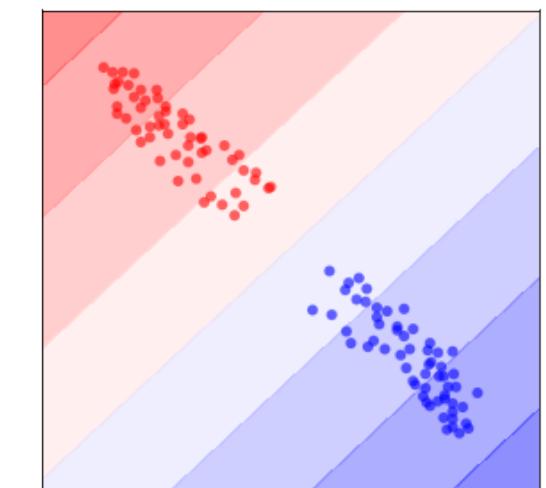
潜在変数



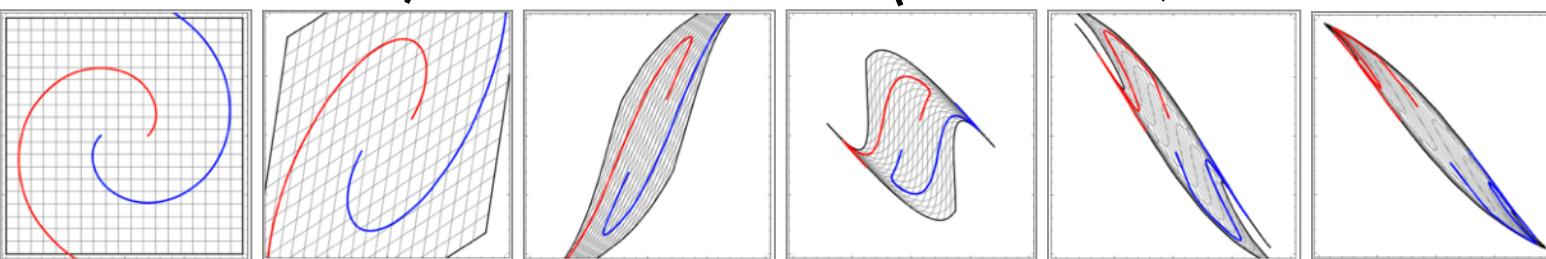
曲面モデル



標準的な
機械学習



x_1

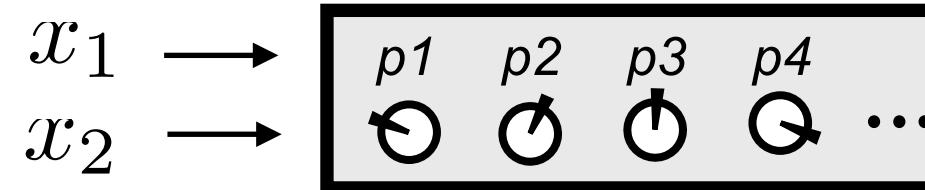


z_1

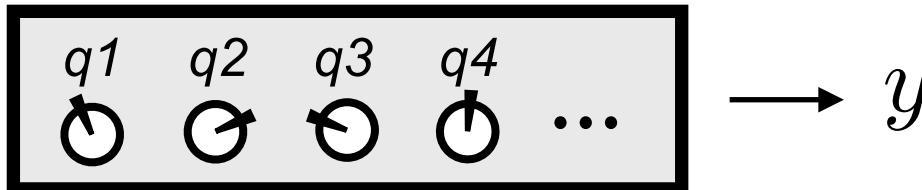
良い変数さえ見つかれば
ここはシンプルで良い！

深層学習と表現学習

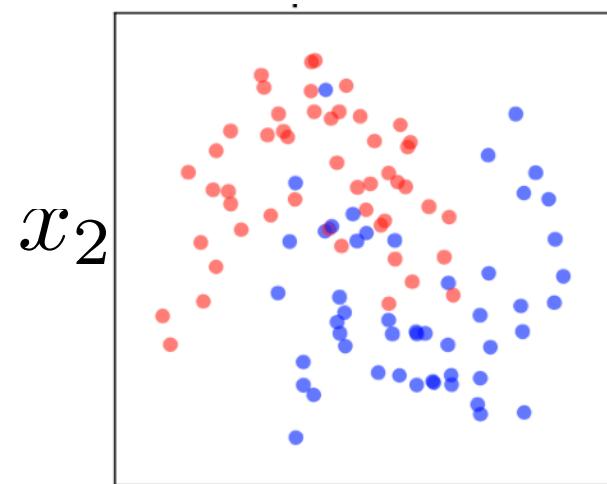
変数変換(表現学習)



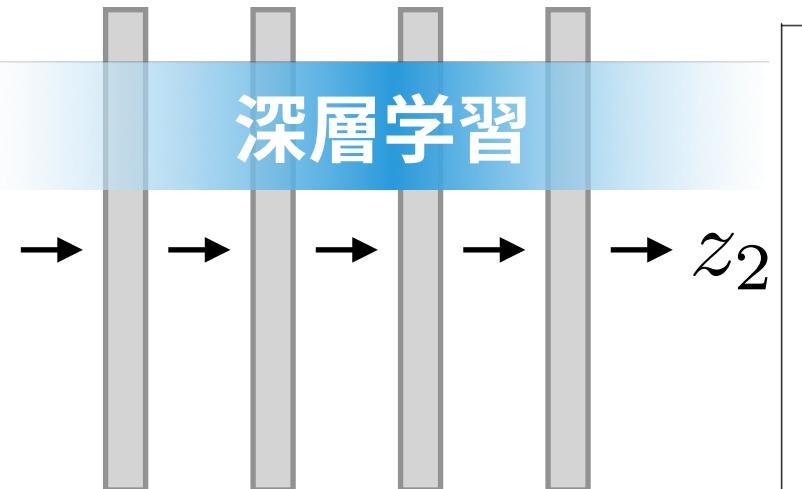
曲面モデル



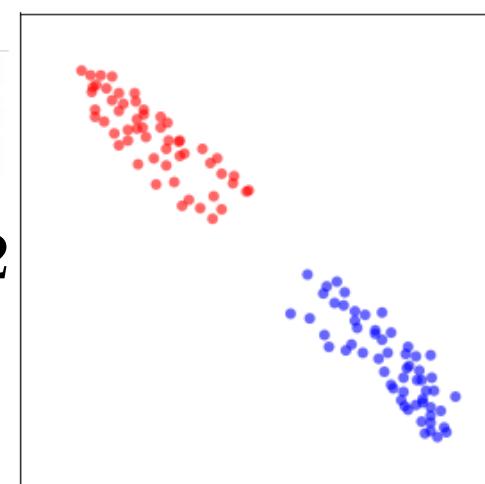
入力変数



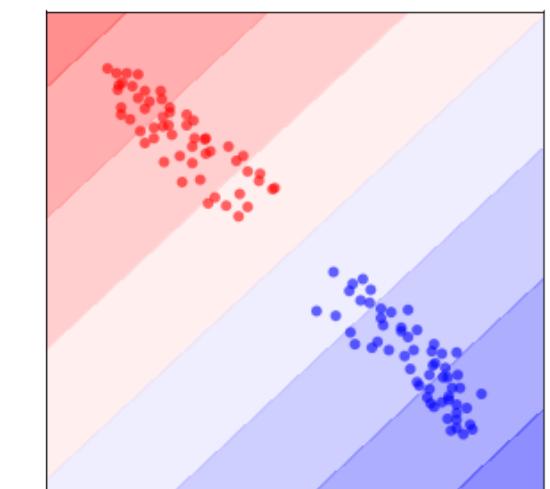
深層学習



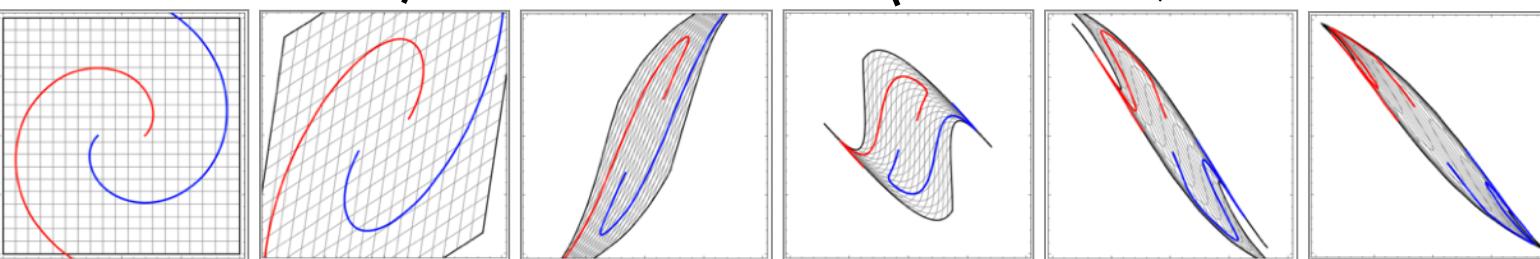
潜在変数



標準的な
機械学習



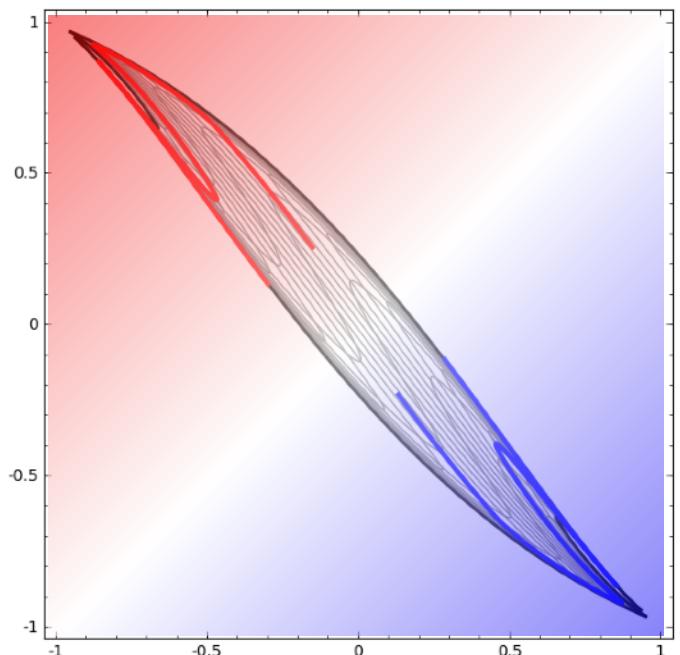
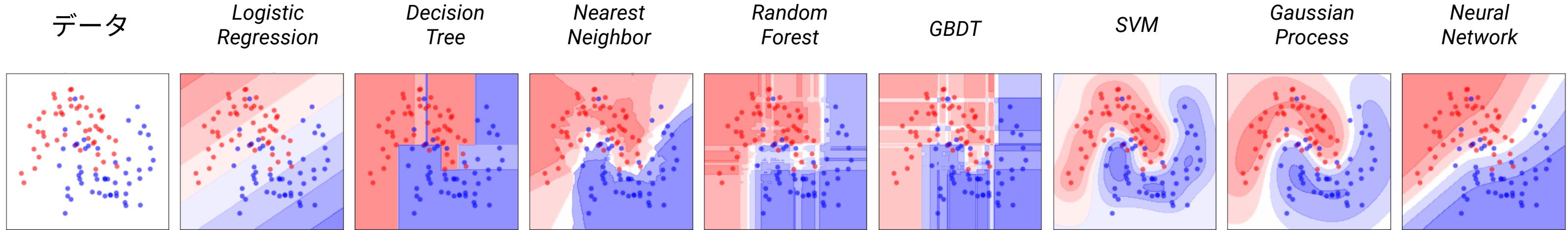
x_1



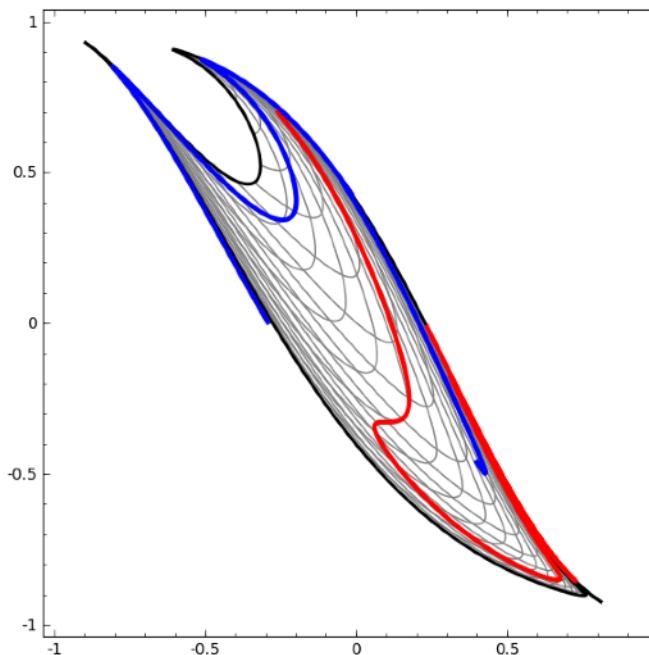
z_1

良い変数さえ見つかれば
ここはシンプルで良い！

深層学習と表現学習



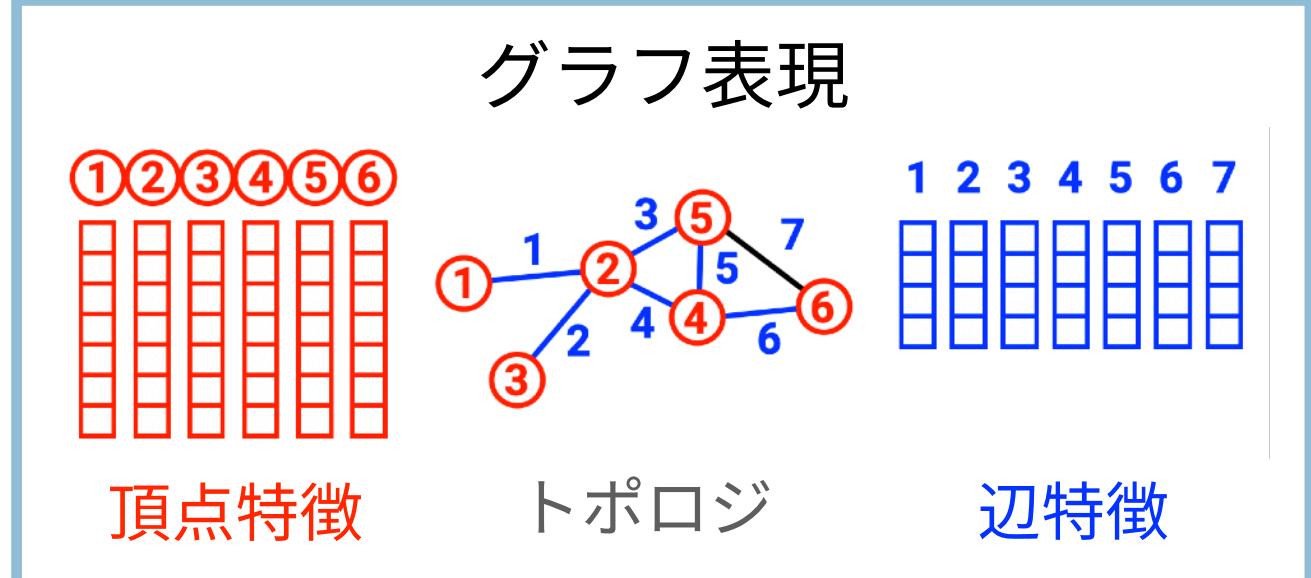
できるだけ線で
分けられるように
変換を学習



いつもうまくいく
とは限らない…。
間違えると元より
難しい問題になる！

実例：分子・材料の表現学習とGraph Neural Networks

分子・材料の環境/条件/標的/相互作用…



表現学習

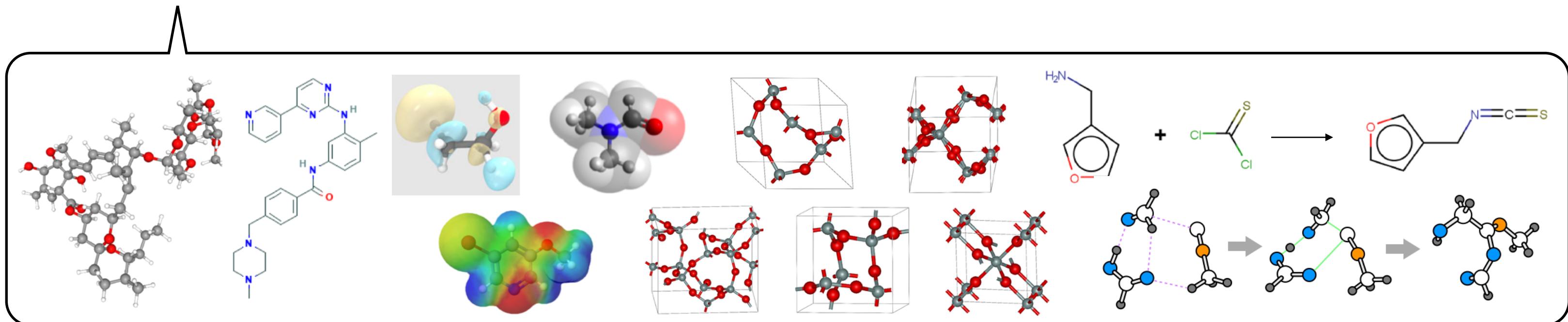
$p_1 \ p_2 \ p_3 \ \dots$

入力情報の合成演算

曲面モデル

$q_1 \ q_2 \ q_3 \ \dots$

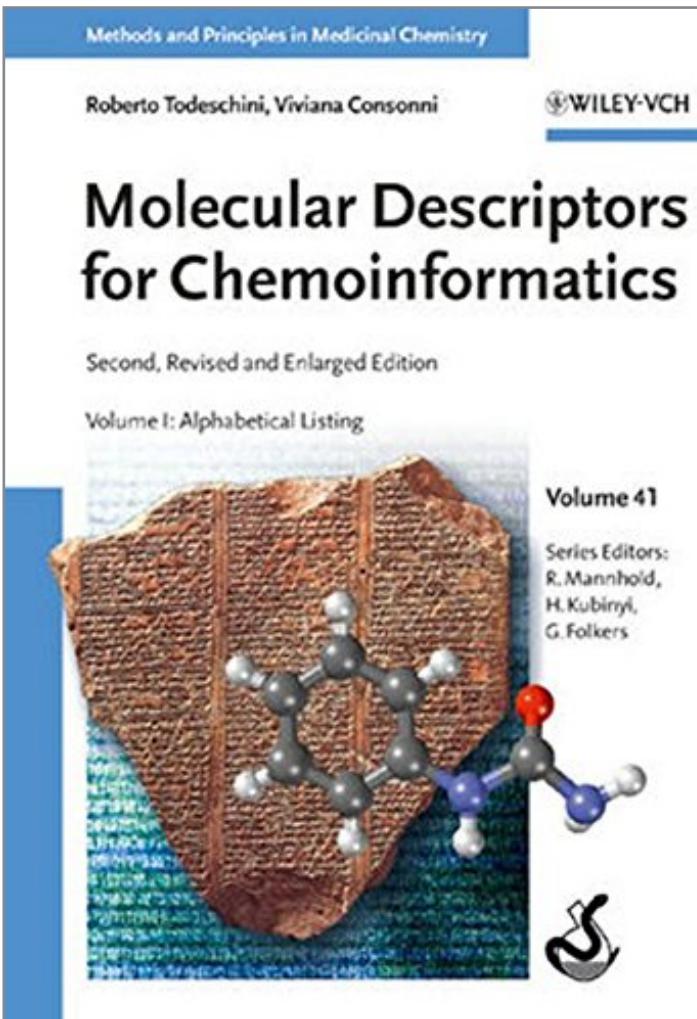
$\rightarrow y$



人間が考えた無数の分子記述子を超える汎用表現はあるか？

- 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, ...

3,300種類の記述子が載っている！



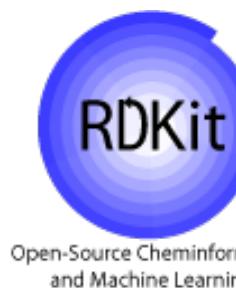
Todeschini and Consonni, *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, 2009.
<https://doi.org/10.1002/9783527628766>

5,270個の記述子を計算してくれる！

DRAGON 7.0



商用の記述子ソフトウェア



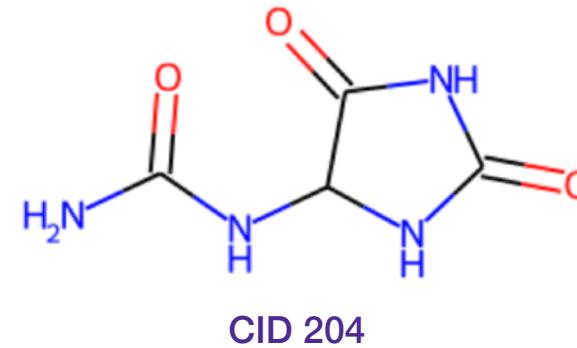
rdkit.Chem

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

rdkit.ML.Descriptors

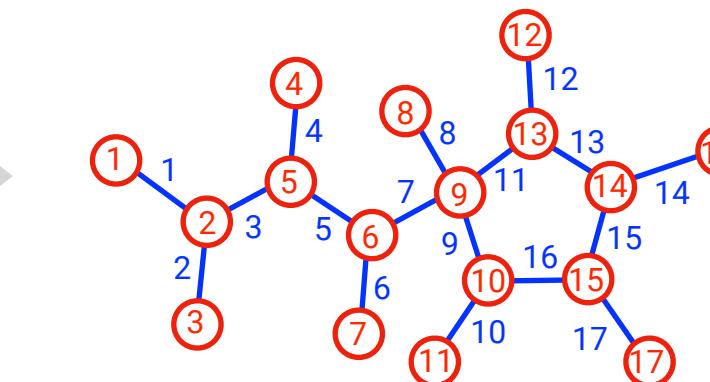
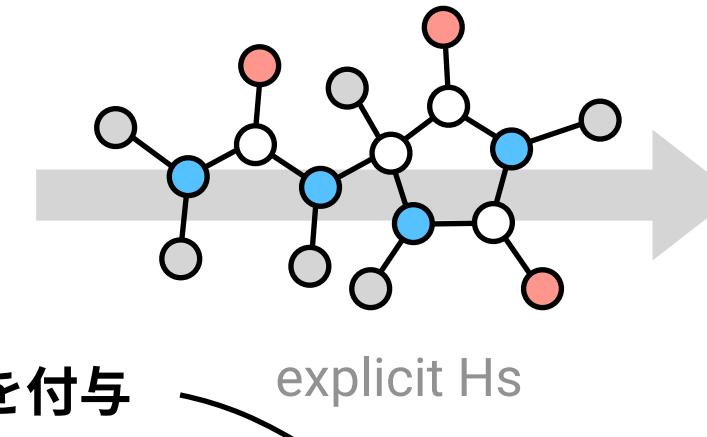
オープンソースフレームワーク

分子のグラフ表現



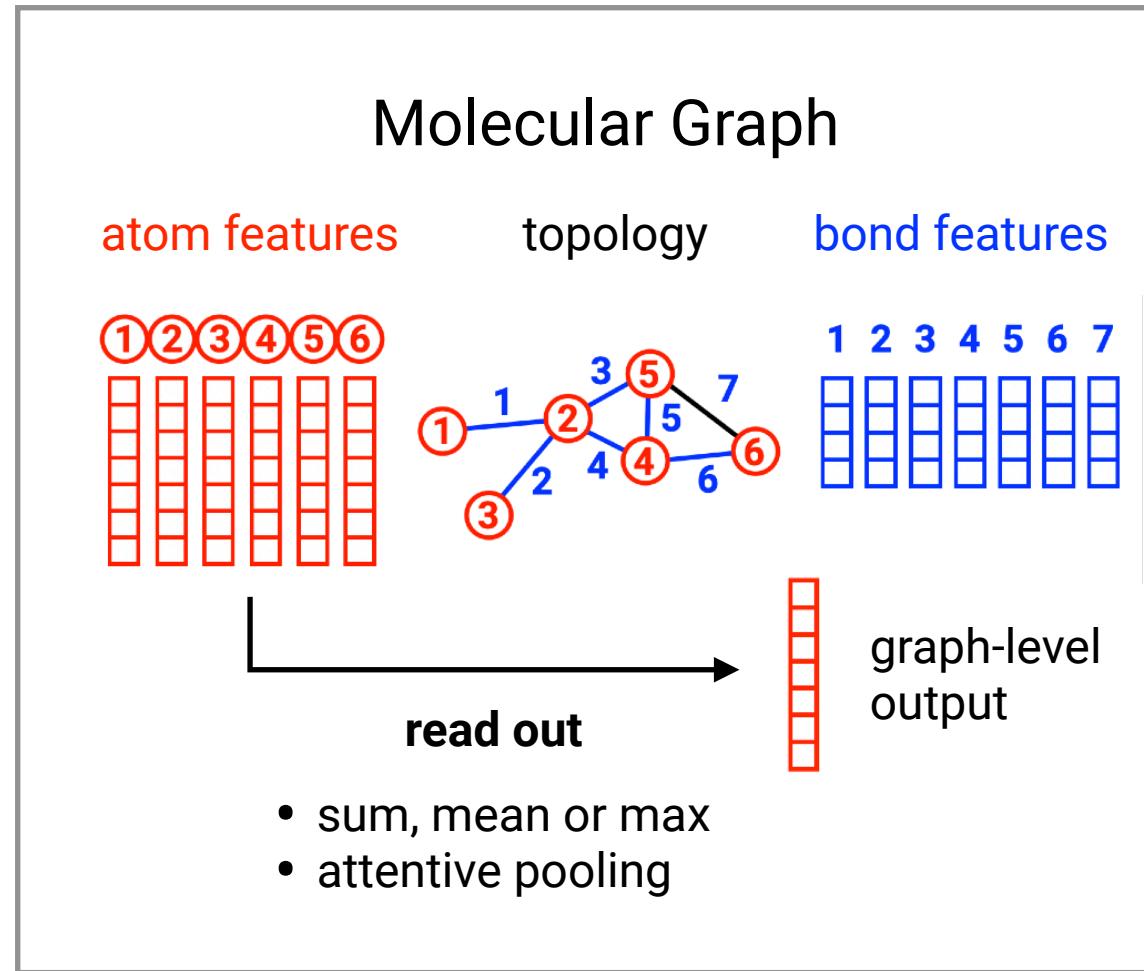
atoms → nodes
bonds → edges
と点と線で抽象化

各点と各線に特徴量を付与



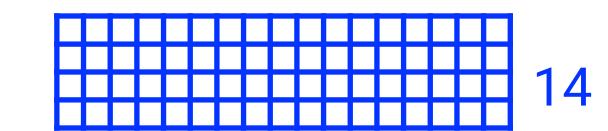
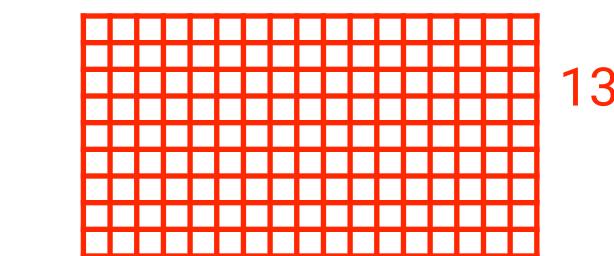
この番号付けの恣意性で
結果が変わらないように

- ! 1. 置換不变性
- 2. 置換同変性



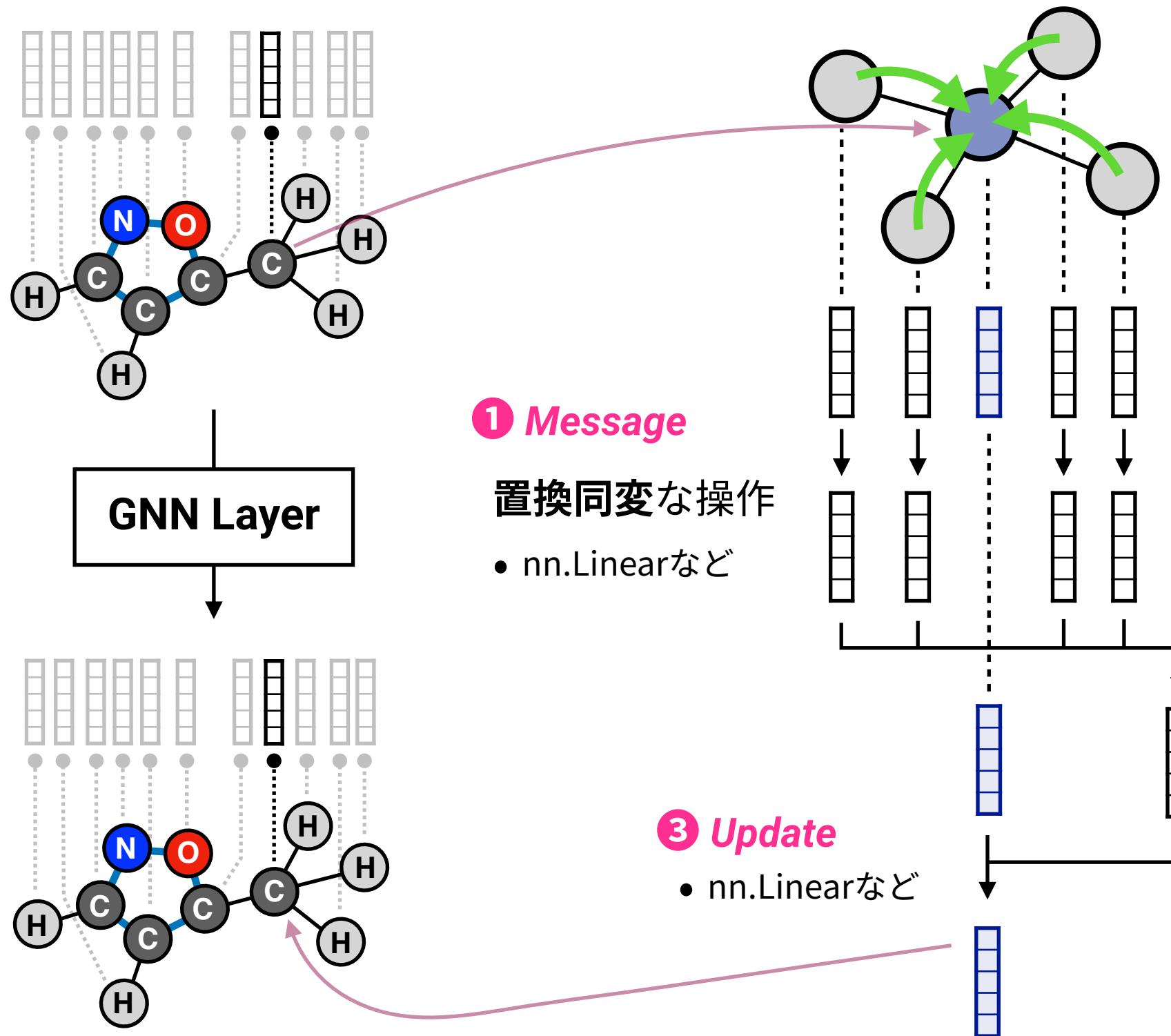
133 features

- 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)



- 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

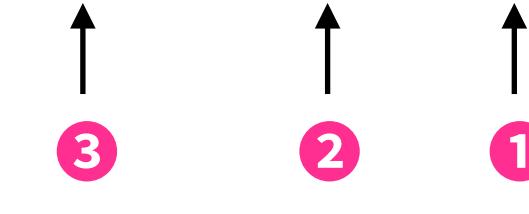


h_i 頂点特徴量

e_{ij} 辺特徴量

近傍からの“Message Passing”で更新

$$h_i \leftarrow \psi \left(h_i, \bigoplus_{j \in \mathcal{N}_i} \phi(h_i, h_j, e_{ij}) \right)$$



辺特徴量は典型的には
この ① で使う

② Aggregate

置換不变な操作

- sum, mean or max
- attentive pooling

有向辺上で特徴量を更新してから頂点特徴量へ合成する
バリエーション(Directed MPNN)はあるがこれが基本形

Virtual Screening (QSAR/QSPR)

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



About Blog Submit Contact

Search PubChem

BIOASSAY RECORD

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

Cite

Download



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

PubChem AID	1
Source	DTP/NCI
External ID	NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Cell Lung cell line
BioAssay Type	Confirmatory
Tested Substances	All (53,554) Active (3,025) Inactive (50,655) Data Table
Tested Compounds	All (51,583) Active (2,814) Inactive (48,922)
Version	2.1 Revision History



Virtual Screening (QSAR/QSPR)

input



CID 11978790

ML

output

activity: “Active” (分類)

LogGI50: -7.8811 (回帰)

GI50: concentration required
for 50% inhibition of growth

Tested Compounds

All (51,583)

Active (2,814)

Inactive (48,922)

Tested Substance			Sort By: Activity				
Structure	CID	SID	Activity	Score	LogGI50_M	LogGI50_u	LogGI50_V
	5298	121832	Active	67	-8		
	363173	493713	Active	43	-6.5871		
	399631	530868	Active	51	-7.0678		
	399630	530867	Active	60	-7.617		

Tested Substance			Sort By: Activity				
Structure	CID	SID	Activity	Score	LogGI50_M	LogGI50_u	LogGI50_V
	390324	521601	Inactive	0	-4		
	390311	521588	Inactive	0	-4		
	390312	521589	Inactive	4	-4.214		
	135489876	521590	Inactive	13	-4.7552		

Virtual Screening (QSAR/QSPR)

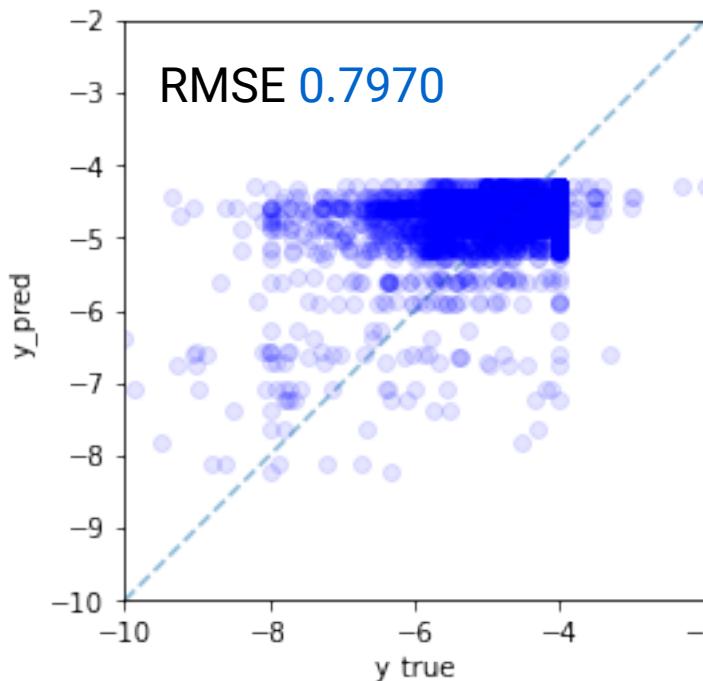
Standard ML

ExtraTrees
w/ ECFP6(1024)

- 予測精度 (分類)

95.079% (Active/Inactive)

- 予測精度 (回帰)



GNN

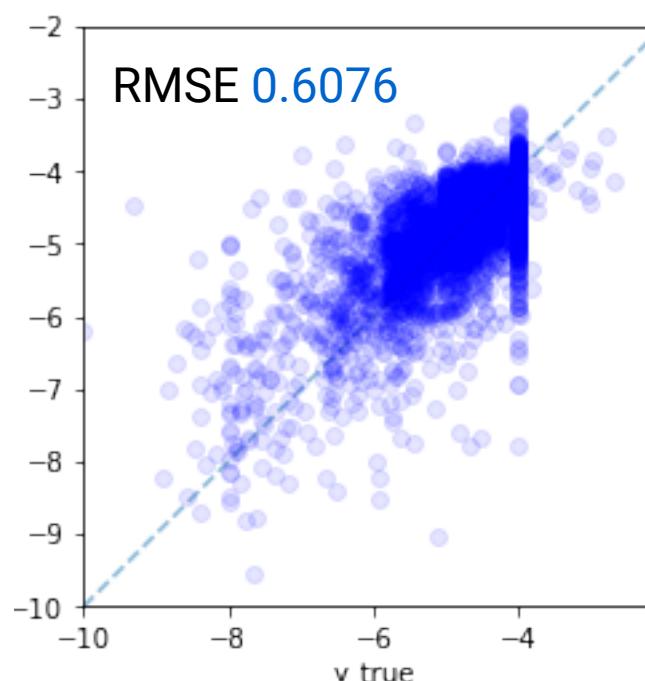
ChemProp
(Directed MPNN)

边上で更新する
シンプルなGNN

- 予測精度 (分類)

95.604% (Active/Inactive)

- 予測精度 (回帰)



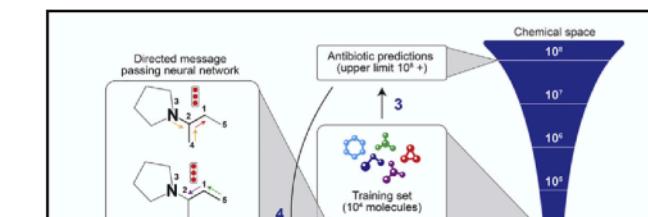
ChemProp (Yang et al, 2019)

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.),
jimjc@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

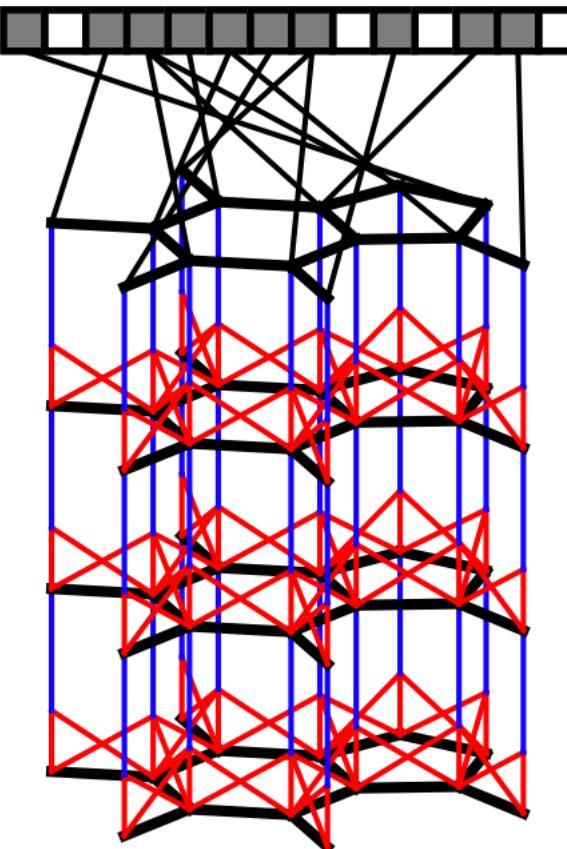
Machine learning spots molecules that work even against 'untreatable' strains of bacteria.

Jo Marchant

Marchant, 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
- 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  $\mathbf{f}$ 
```

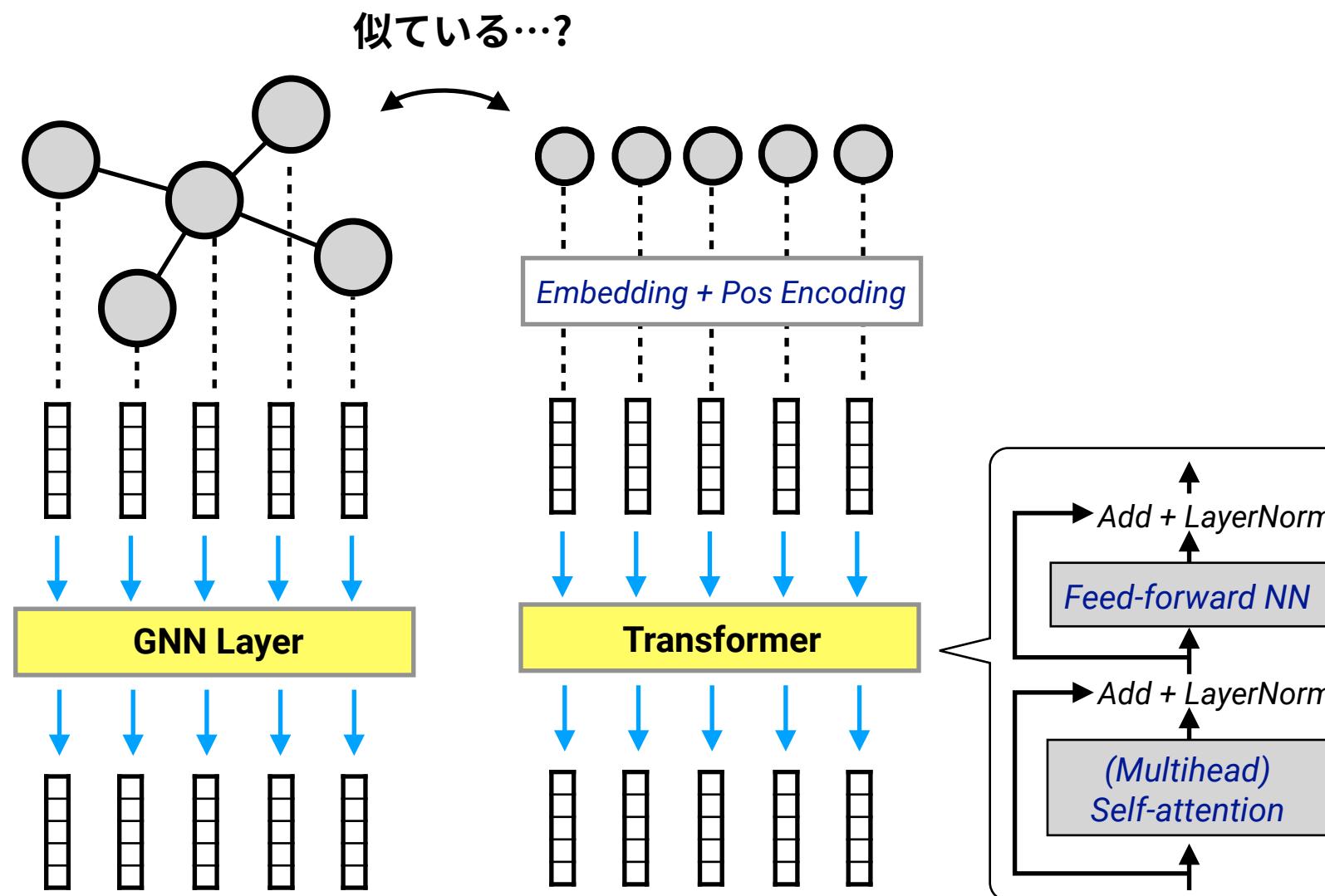
Algorithm 2 Neural graph fingerprints

```
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:       $i \leftarrow \text{softmax}(\mathbf{r}_a W_L)$          $\triangleright$  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
- 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>

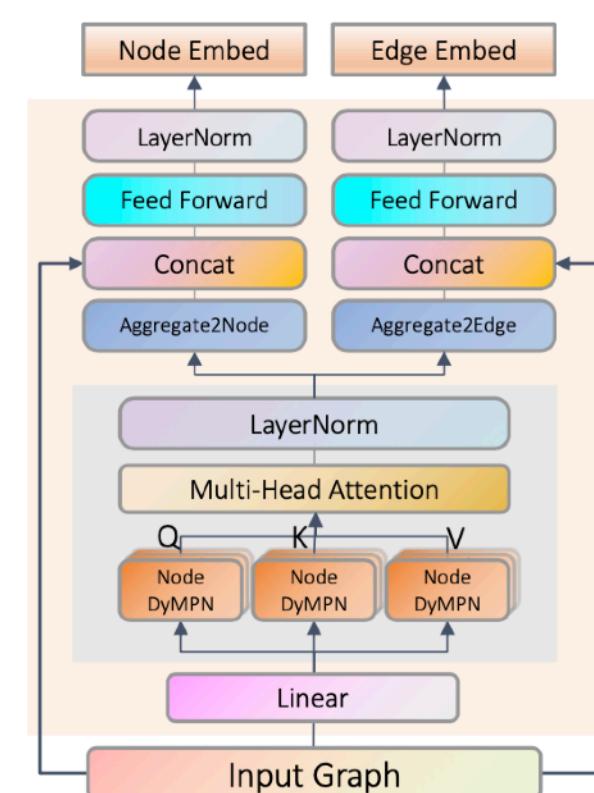
↓
Ying et al (2021) のGraphomerはKDDCup 2021の
Graph-levelタスクの優勝モデルで使われた

事前学習が効かないとされたNLPや
CNN一択かに見えたCVを変革してきた
Transformerは分子タスクもえるのか？

小サンプル問題の克服：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
- 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や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>



[Strategies for Pre-training Graph Neural Networks](#)
Hu, Liu, Gomes, Zitnik, Liang, Pande, Leskovec (ICLR 2020)
<https://arxiv.org/abs/1905.12265>

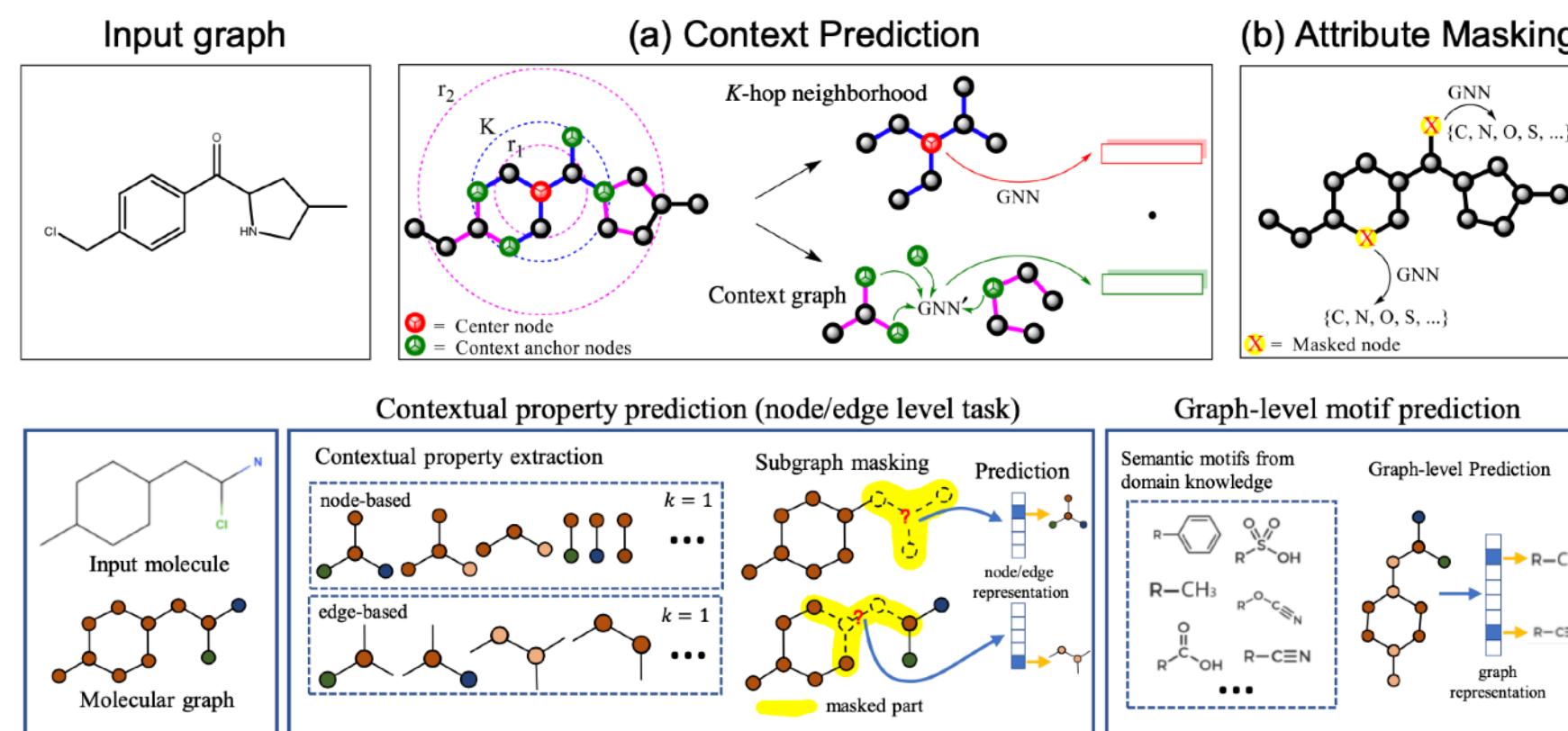


Figure 1: Overview of GTransformer.

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

Self-Supervisedなので
大規模データがあれば
教師ラベル付けが不要

事前学習した分子表現は
様々な下流タスクで転移

- 多様な分類・回帰
- 分子生成

分子表現の生成：Generative Chemistry

- 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
- Decoderが非自明、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也要検証

Deep Graph Generators: A Survey

FAEZEH FAEZ¹, YASSAMAN OMMI², MAHDIEH SOLEYMANI BAGHSHAH¹, AND HAMID R. RABIEE¹, (Senior Member, IEEE)

¹Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

²Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran

Corresponding authors: Hamid R. Rabiee and Mahdieh 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]

OGB Large-Scale Challenge (KDDCup2021)

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

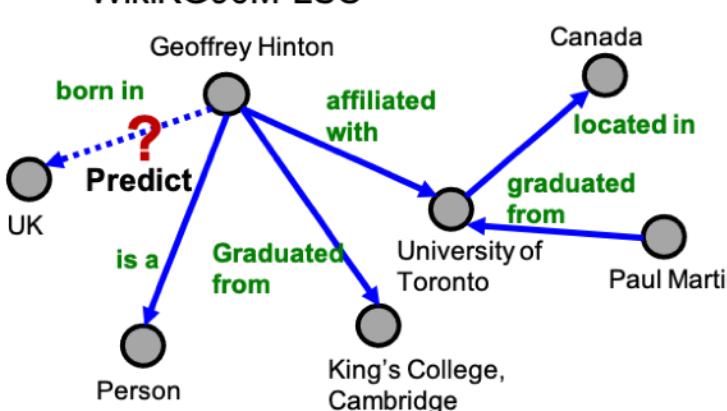
Node-level

MAG240M-LSC



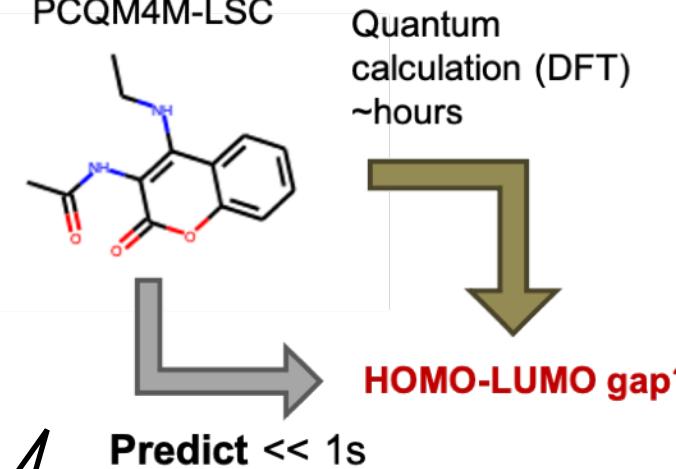
Link-level

WikiKG90M-LSC



Graph-level

PCQM4M-LSC



OPEN GRAPH BENCHMARK

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

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)**

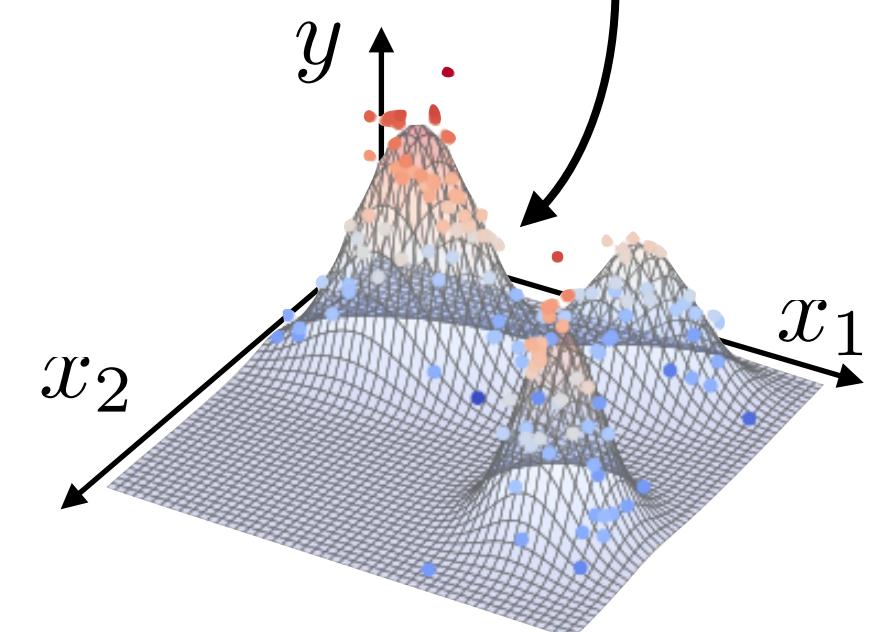
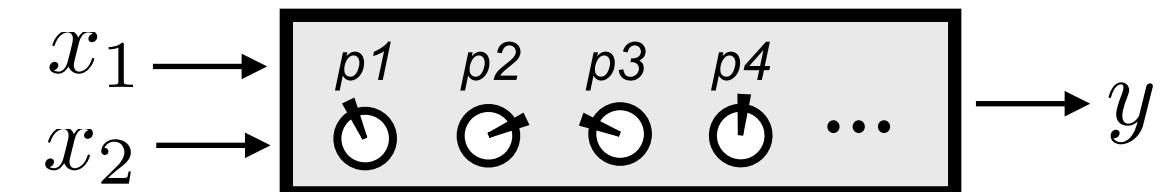
機械学習の現代的側面① 入力変数がめっちゃ多い…

右の絵を眺めると、機械学習が x_1 と x_2 以外の
入力されてない情報を全く考慮してくれないことは一目瞭然

→ 出力の予測に本当は必要な情報を入力していなかつたら
機械学習は擬似相関に過ぎず何も本質を捉えられない

spurious correlation

曲面モデル



機械学習の現代的側面① 入力変数がめっちゃ多い…

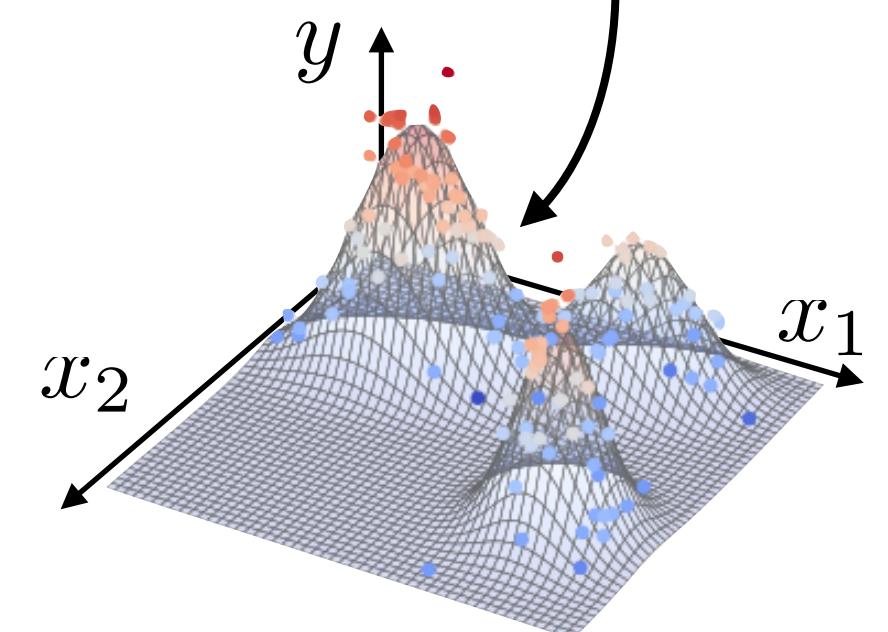
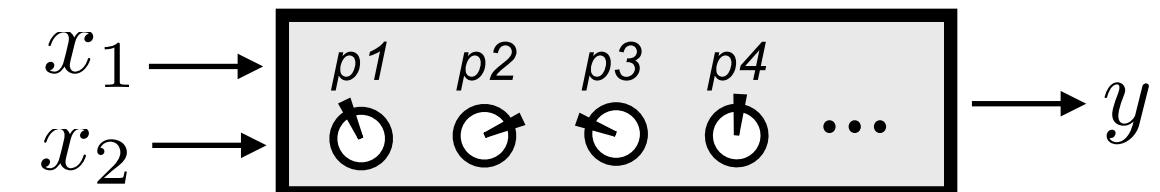
右の絵を眺めると、機械学習が x_1 と x_2 以外の
入力されてない情報を全く考慮してくれないことは一目瞭然

→ 出力の予測に本当は必要な情報を入力していなかつたら
機械学習は擬似相関に過ぎず何も本質を捉えられない
spurious correlation

機械学習×化学：スタートラインでのつまづき

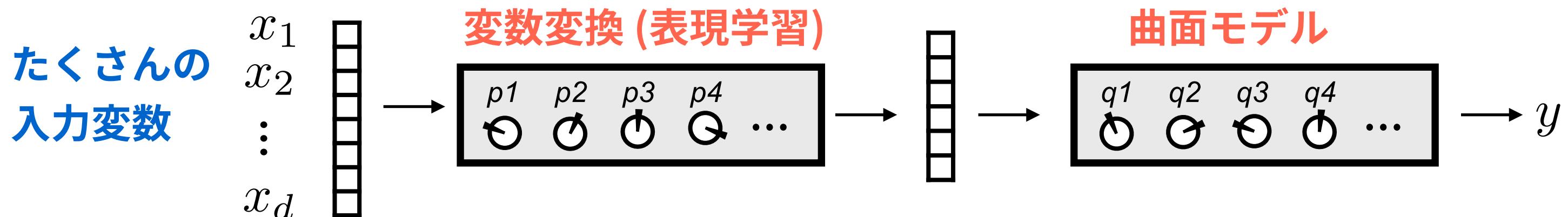
- 多くの場合、関心の出力を得るために必要十分な入力が何
なのかはよく知らない
- というか、そもそもよく分からないから機械学習を使いたい。
なのに、機械学習がうまくいくためにはどんな入力が必要か
理解しておく必要があるってどゆこと！？ 😠

曲面モデル



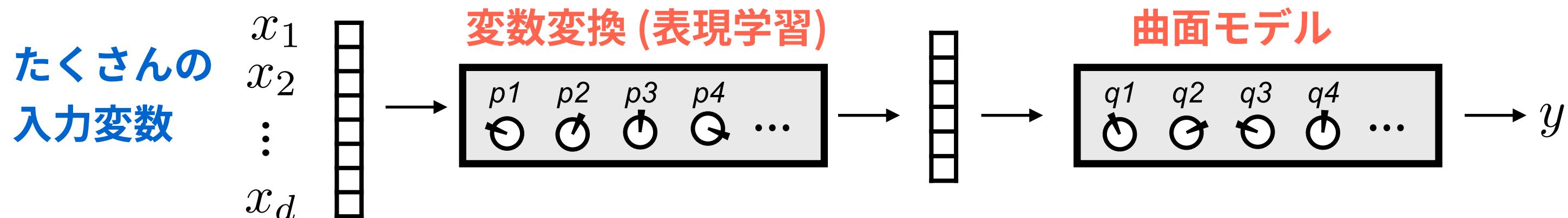
機械学習の現代的側面① 入力変数がめっちゃ多い…

必然的に「少しでも関係ありそうな情報は入力して」表現学習で重要なものを峻別という戦略に

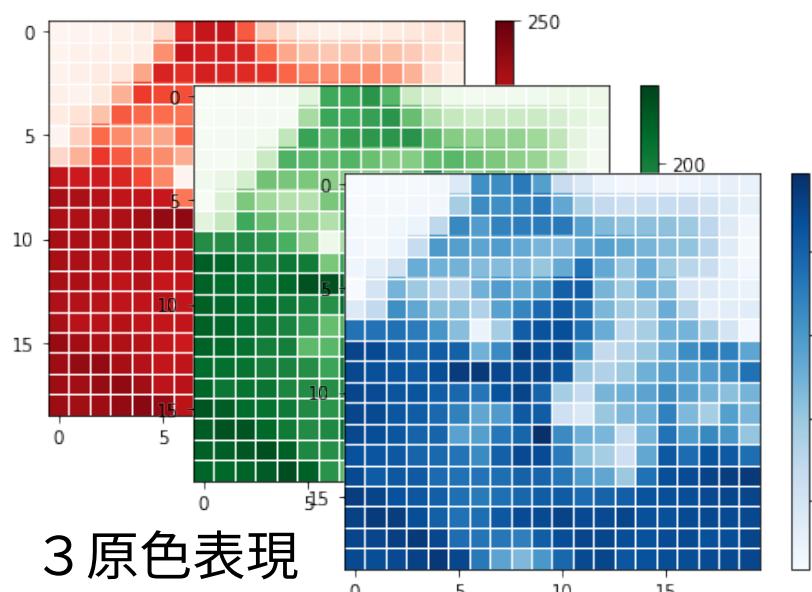
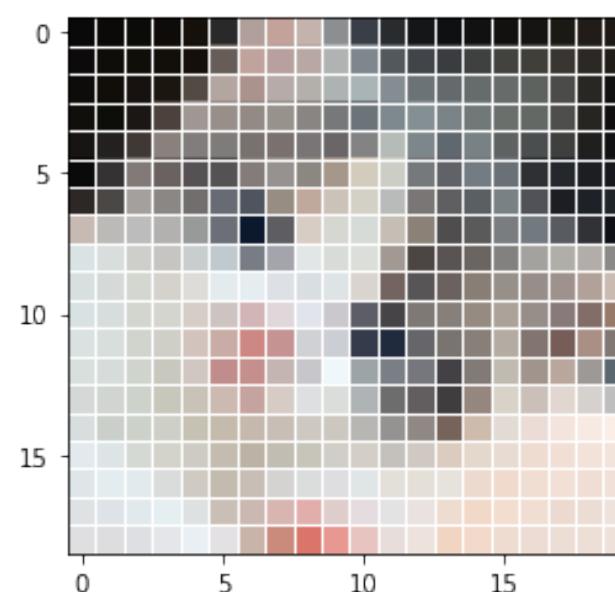


機械学習の現代的側面① 入力変数がめっちゃ多い…

必然的に「少しでも関係ありそうな情報は入力して」表現学習で重要なものを峻別という戦略に



例：「画像に何が写っているか」なら画像の各ピクセルの輝度値を全部そのまま入れる！

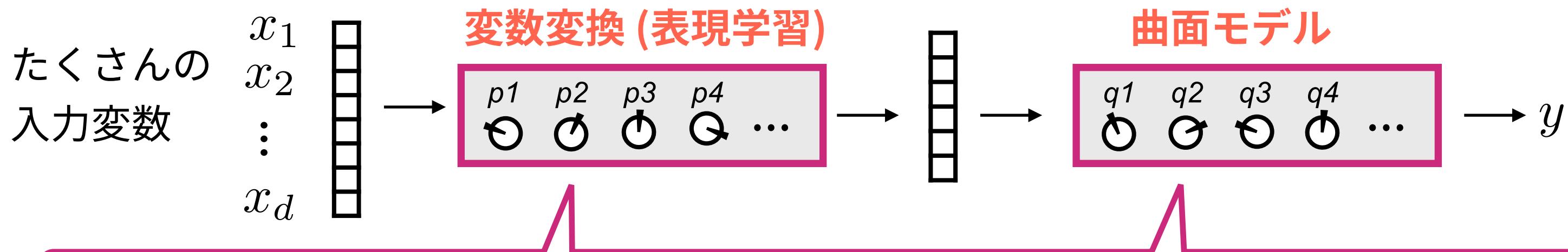


20×20ピクセルのカラー画像は
20×20×3 = 1200個の数値のあつまり

20×20 → 1,200変数
1000×1000 → 300,000変数

機械学習の現代的側面② 内部パラメタの数も死ぬほど多い…

高次元の入出力関係がどのようなものであっても表現するためパラメタ数も死ぬほど多い！



ResNet50: 2600万パラメタ

ResNet101: 4500万パラメタ

EfficientNet-B7: 6600万パラメタ

VGG19: 1億4400万パラメタ

12-layer, 12-heads BERT: 1億1000万パラメタ

24-layer, 16-heads BERT: 3億3600万パラメタ

GPT-2 XL: 15億5800万パラメタ

GPT-3: 1750億パラメタ

現代の機械学習の技術研究が向き合う設定：

1750億個のパラメタ値を持つモデルを数十万の変数を持つ数千万個のデータにフィッティング

結果、現在の機械学習モデルはアホみたいにデータを食う…

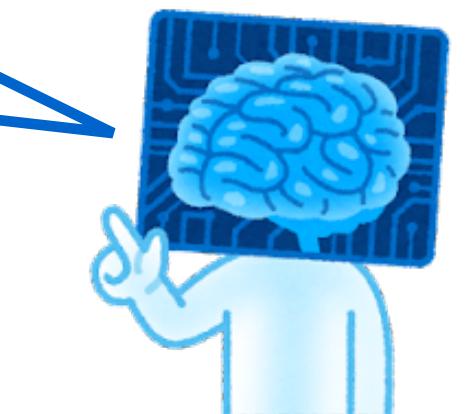
現代の機械学習は強力で良いデータが十分にあれば複雑な入出力関係でも学習できる！

↑
↓ 機械学習×化学：理論・潜在的 possibility と現実の大きなギャップ

現実には、解空間の広さから考えれば「ビッグデータ」ですら「十分」の水準にはほど遠い…

ショボい認知能力のおまえら人間にとったら「ビッグ」データかもしらんけど、
ホンマに必要な情報量からしたらハナクソみたいなもんやな！

by ディープラーニング様



- 何でもかんでも入力変数に入れまくる全部入りモデルは「キッチンシンク回帰」と揶揄され
伝統的な応用統計学ではタブーだった…（過適合のリスクが大きすぎて良いことないから）
- "良性の"過適合："ビッグデータ"では過適合 자체がそもそも難しいので気にしない立場も

羅生門効果とUnderspecification

羅生門効果：良い機械学習モデルの多重性（非一意性）

同じ見本例データから同程度の高い予測精度を持つ良い機械学習モデルは無数に作れる！

原因：入力変数の選び方、モデルの選択や設計、初期値の違い、パラメタの多重性、…

羅生門効果とUnderspecification

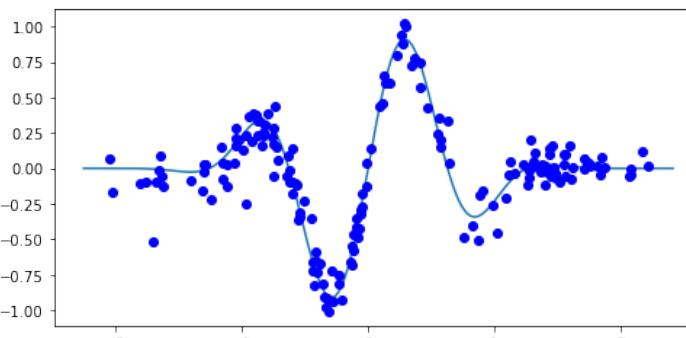
羅生門効果：良い機械学習モデルの多重性（非一意性）

同じ見本例データから同程度の高い予測精度を持つ良い機械学習モデルは無数に作れる！

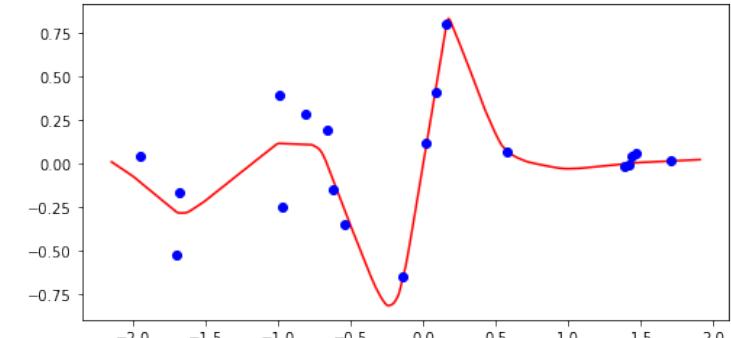
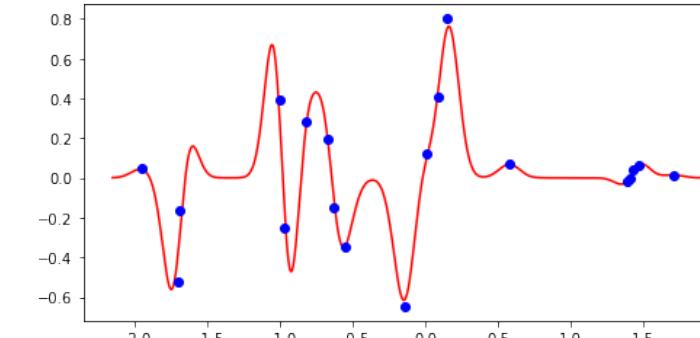
原因：入力変数の選び方、モデルの選択や設計、初期値の違い、パラメタの多重性、…

- 現象の「理解」のため獲得した機械学習モデルを分析して示唆を引き出そうとするときの最大の障害。真実味を帯びた解釈が無数にあることになりまさに真実は「藪の中」…
- さらに実際は本質的にデータが足りてない(Underspecification)ことで多重性はさらに悪化

だいたいの方法で類似



手法やモデルによって予測時の挙動にかなり差が出てしまう



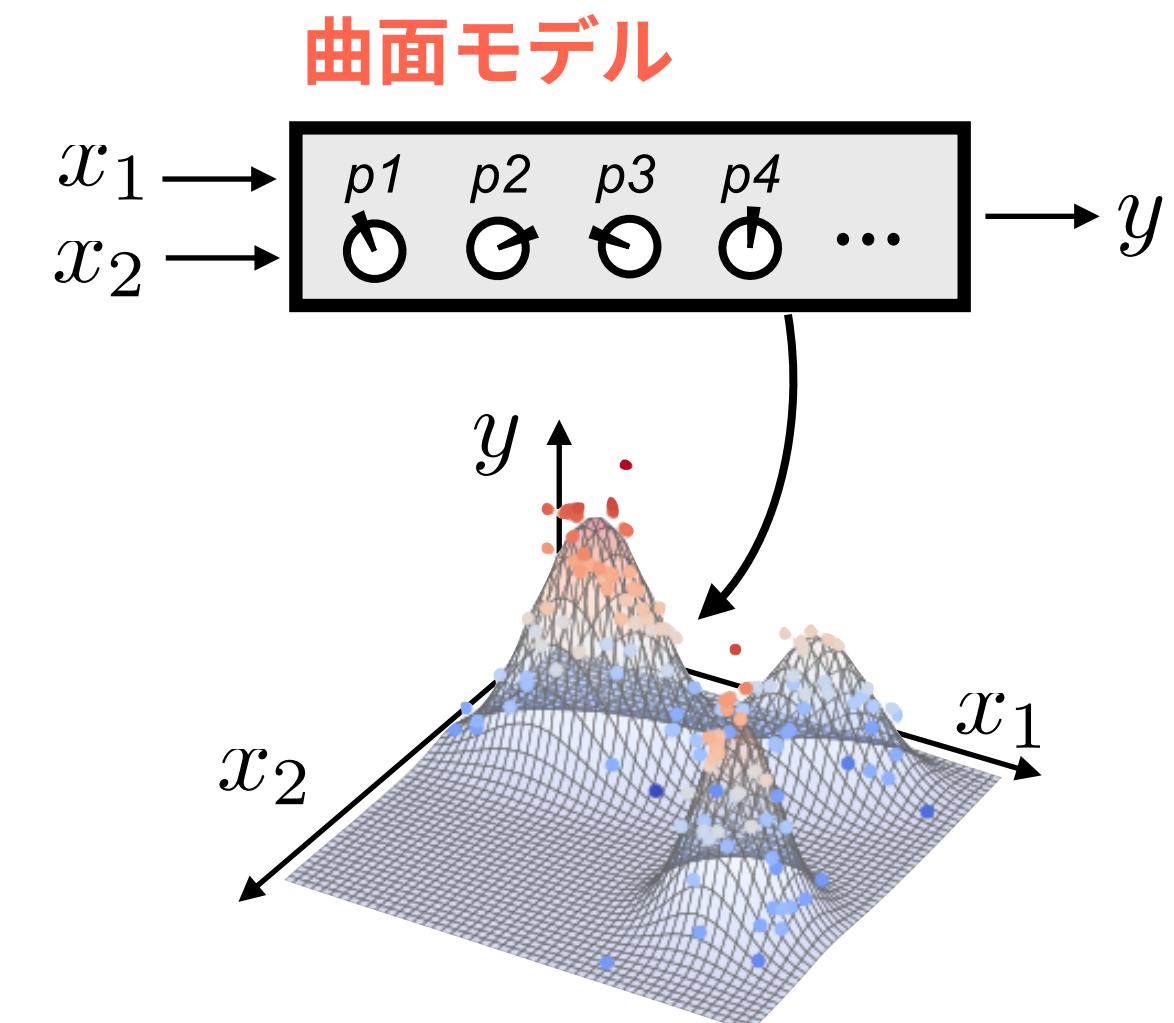
現代の技術的関心はこの高次元性をどう手懐けるか

1. 確率的最適化・正則化 → モデルが大きい自由度の中で暴れまくらないよう動ける範囲を何とかして制御・制限・安定化する
2. 事前学習 (Warm Start) の転移 → 事前に探しておいた良い感じのパラメタ初期値を使う

現代の技術的関心はこの高次元性をどう手懐けるか

1. 確率的最適化・正則化 → モデルが大きい自由度の中で暴れまくらないよう動ける範囲を何とかして制御・制限・安定化する
2. 事前学習 (Warm Start) の転移 → 事前に探しておいた良い感じのパラメタ初期値を使う
3. 帰納バイアスの設計

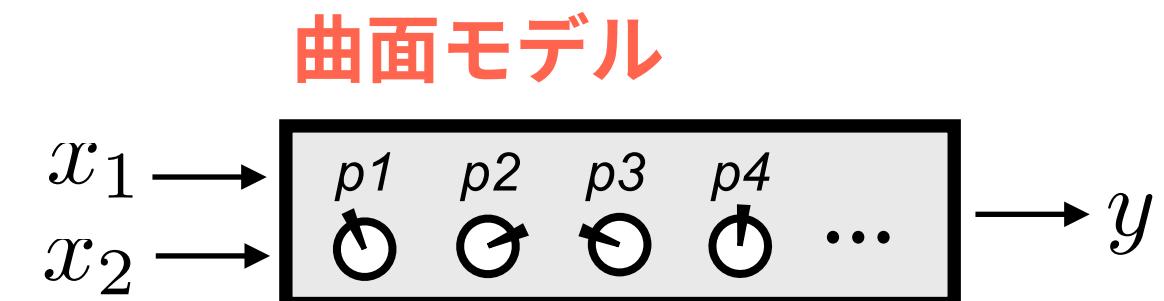
曲面モデルがどんな入出力関係でも表現できることが逆に擬似相関やUnderspecificationの問題を悪化させている



現代の技術的関心はこの高次元性をどう手懐けるか

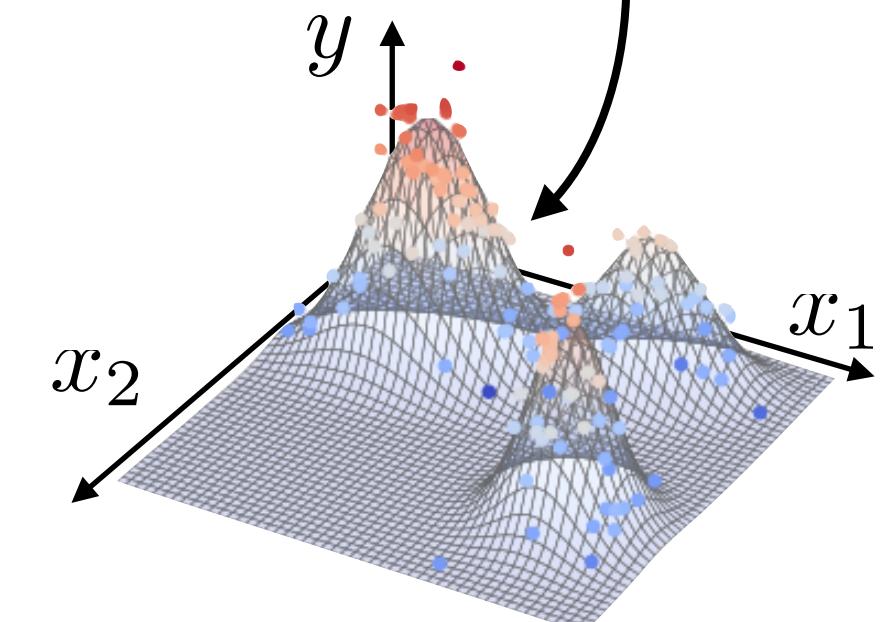
1. 確率的最適化・正則化 → モデルが大きい自由度の中で暴れまくらないよう動ける範囲を何とかして制御・制限・安定化する
2. 事前学習 (Warm Start) の転移 → 事前に探しておいた良い感じのパラメタ初期値を使う
3. 帰納バイアスの設計

曲面モデルがどんな入出力関係でも表現できることが逆に擬似相関やUnderspecificationの問題を悪化させている



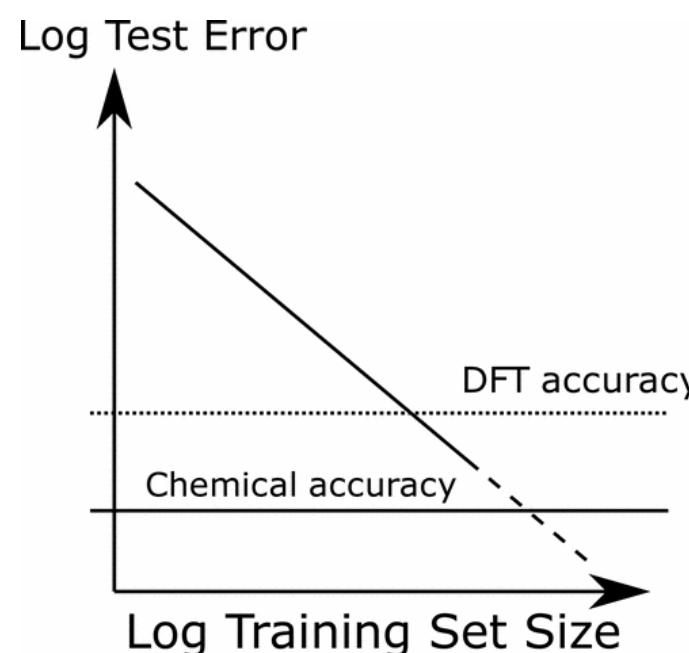
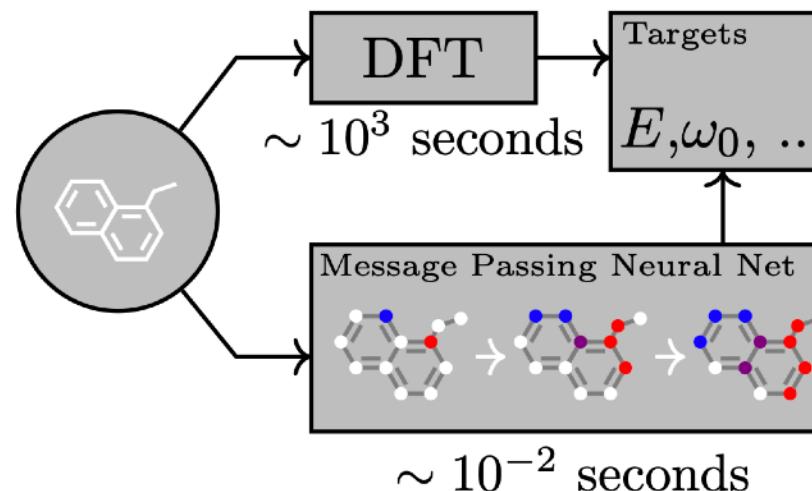
機械学習×化学：化学に適合した帰納バイアスのデザイン

化学的に妥当性を欠くようなモデルが意図せず表現されてしまわないように化学の知識や理論科学・計算化学の知見を総動員して**モデルの自由度を技術的に制限**する！



グレイボックス最適化：化学に適合した帰納バイアスの設計

Googleが色々なGNNをMPNN(Message Passing NN)として統一化した際、DFT計算近似をターゲットに



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

Neural Message Passing for Quantum Chemistry

Justin Gilmer¹ Samuel S. Schoenholz¹ Patrick F. Riley² Oriol Vinyals³ George E. Dahl¹

機械学習の会議で
発表された

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,[†] Luke Hutchison,[‡] Bing Huang,[†] Justin Gilmer,[†] Samuel S. Schoenholz,[†] George E. Dahl,[†] Oriol Vinyals,[§] Steven Kearnes,[†] Patrick F. Riley,[†] and O. Anatole von Lilienfeld^{*,†,ID}

グレイボックス最適化：化学に適合した帰納バイアスの設計

ケモインフォマティクスの原子不变量

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

ECFPの原子不变量

- 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

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?

ECFPのアルゴリズムでは
原子不变量に連續値を入れるのは
ありえなかった(頂点ラベルの役割)

MPNNで用いられた頂点・辺特徴

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

Table 1. Atom Features for the MG Representation^a

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

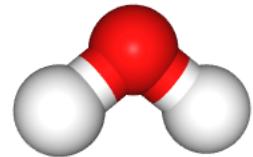
Table 2. Atom Pair Features for the MG Representation^a

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

SchNet: 幾何的GNNの先駆的スタンダード

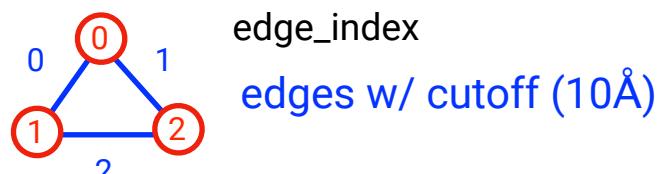
input molecule H₂O



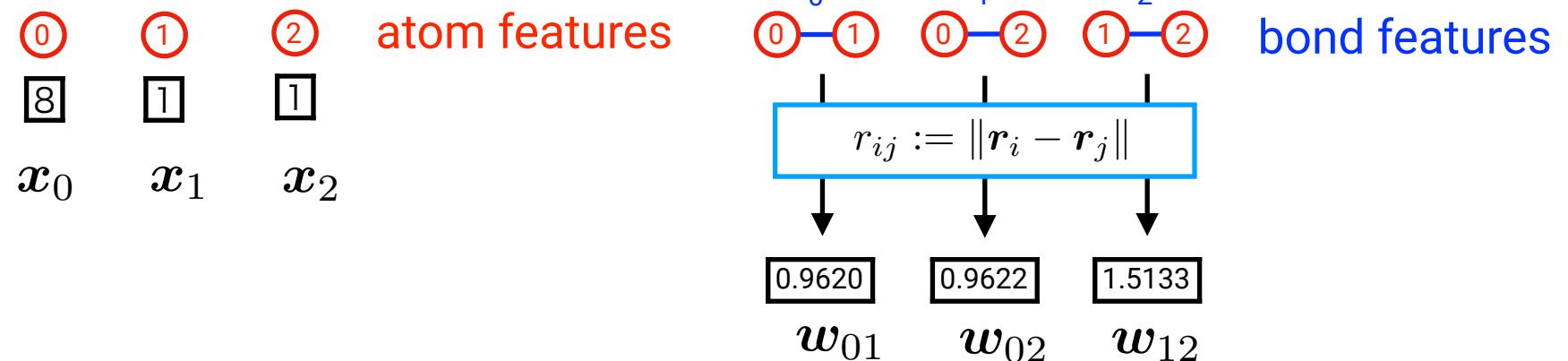
gdb_3

	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

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 (Schütt et al, 2017)

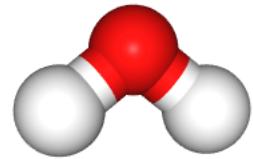
Message Passing with
residual connections

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

Bio-QSARに使うGNNと比べるとかなり
帰納バイアスを**化学に寄せて**設計

SchNet: 幾何的GNNの先駆的スタンダード

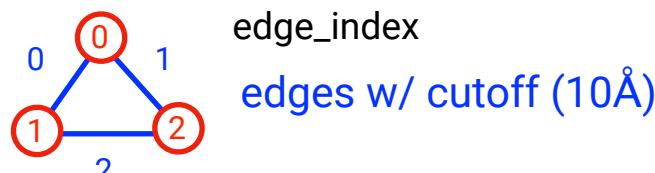
input molecule H₂O



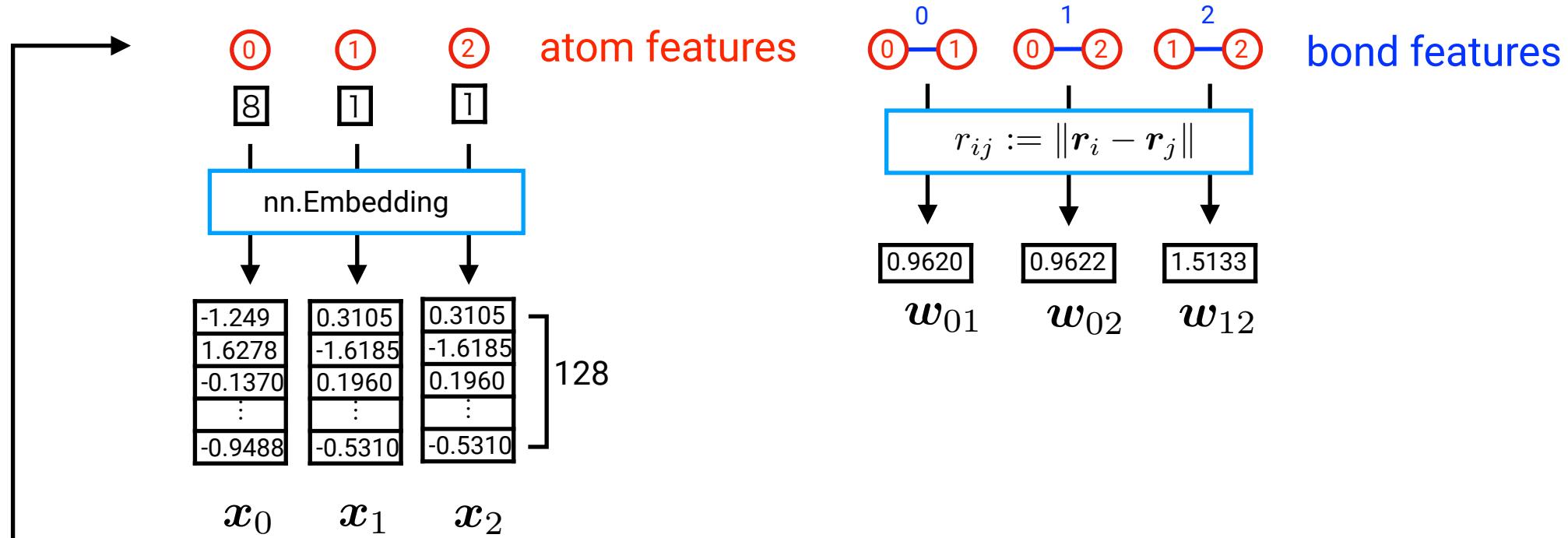
gdb_3

	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

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 (Schütt et al, 2017)

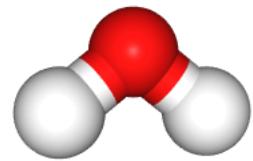
Message Passing with residual connections

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

Bio-QSARに使うGNNと比べるとかなり
帰納バイアスを化学に寄せて設計

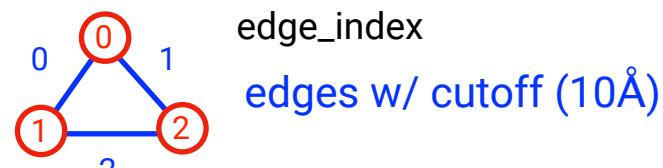
SchNet: 幾何的GNNの先駆的スタンダード

input molecule H₂O

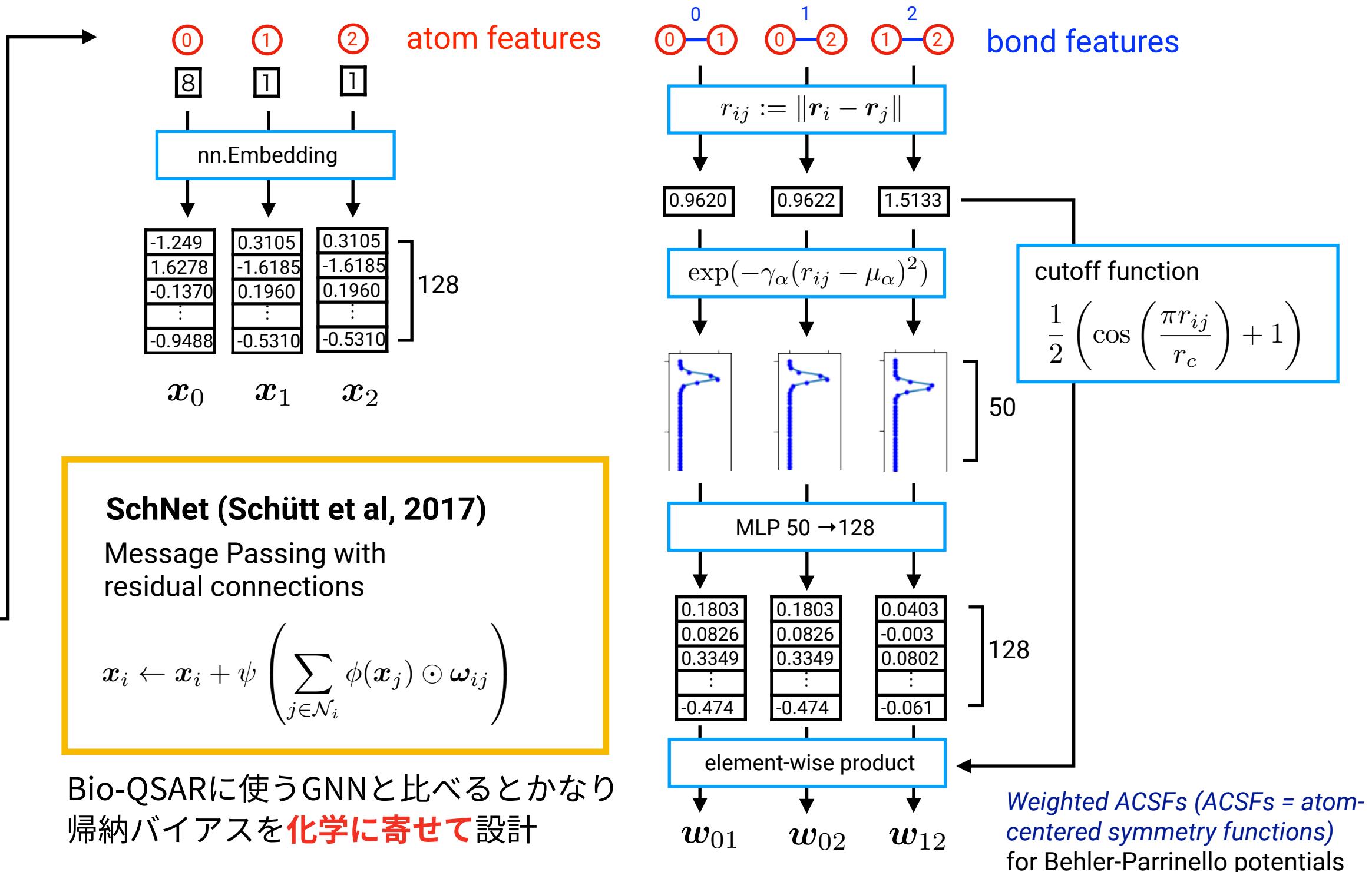


gdb_3				
	x	y	z AN	
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

graph (SchNet)

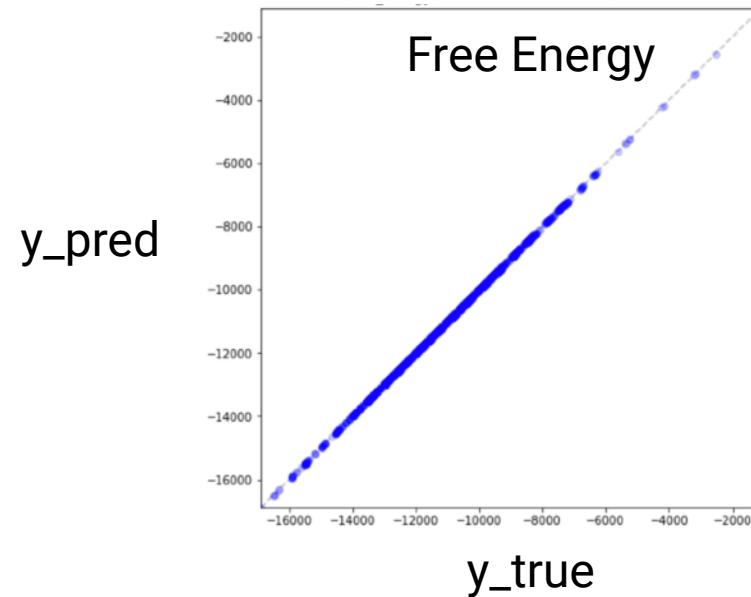


$$\begin{aligned} \mathbf{r}_0 &= [-0.0344, 0.9775, 0.0076] & Z_0 &= 8 \\ \mathbf{r}_1 &= [0.0648, 0.0206, 0.0015] & Z_1 &= 1 \\ \mathbf{r}_2 &= [0.8718, 1.3008, 0.0007] & Z_2 &= 1 \end{aligned}$$

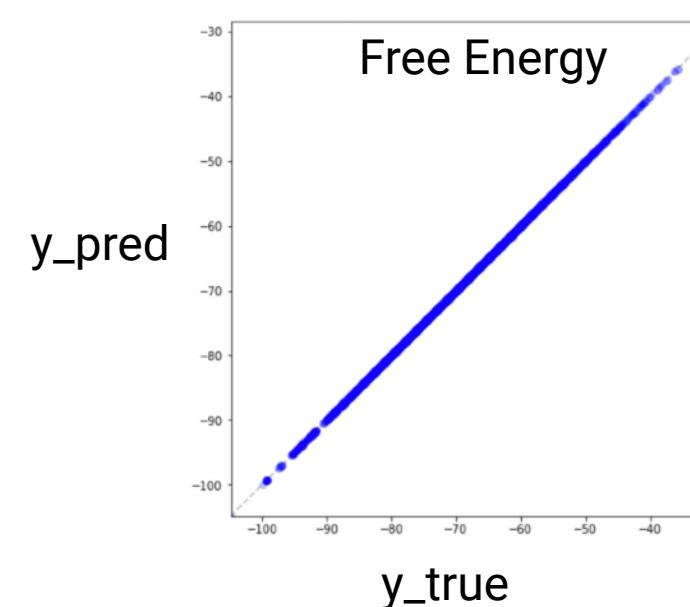


SchNet: 幾何的GNNの先駆的スタンダード

SchNet (Schütt et al, 2017)

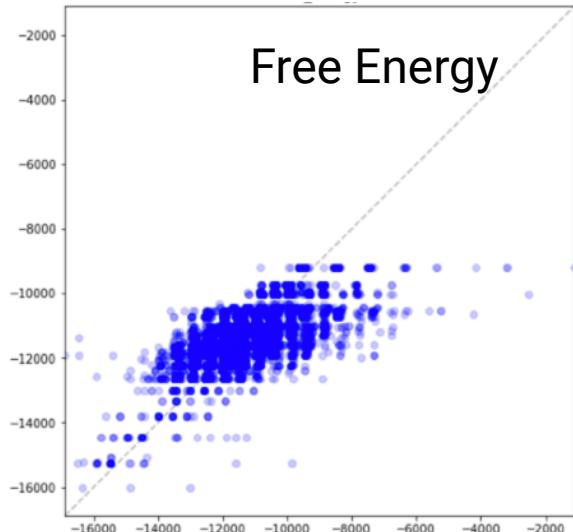


DimeNet (Klicpera et al, 2020)

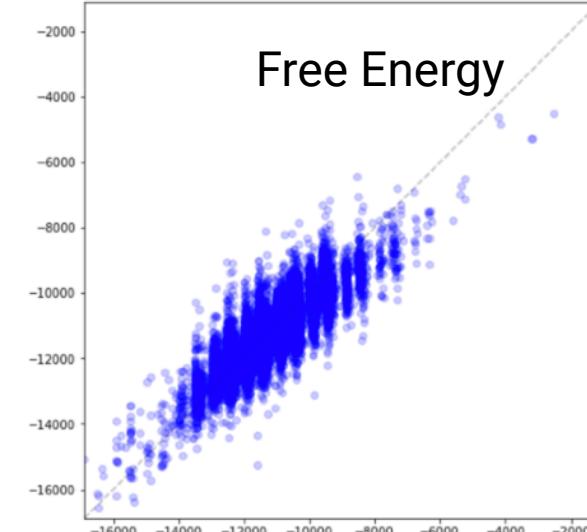


結合角を考慮し
辺上で更新する
改良版
(力も予測する)

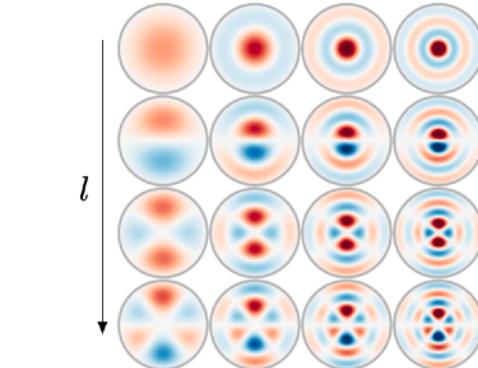
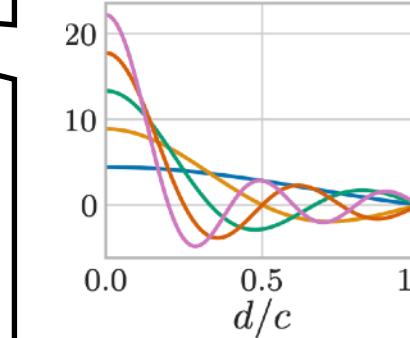
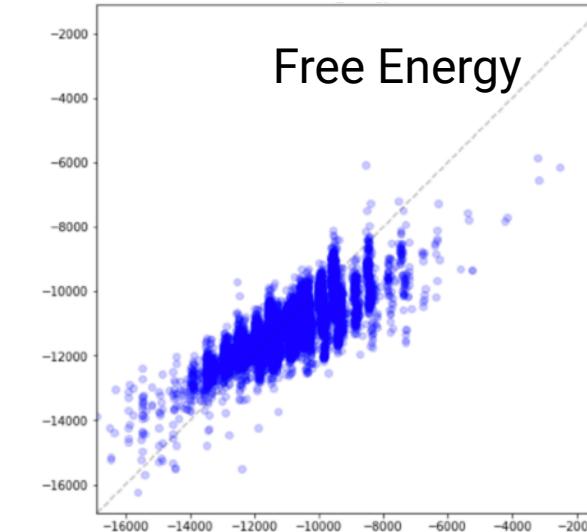
ExtraTrees w/ ECFP6
(without 3D geometry)



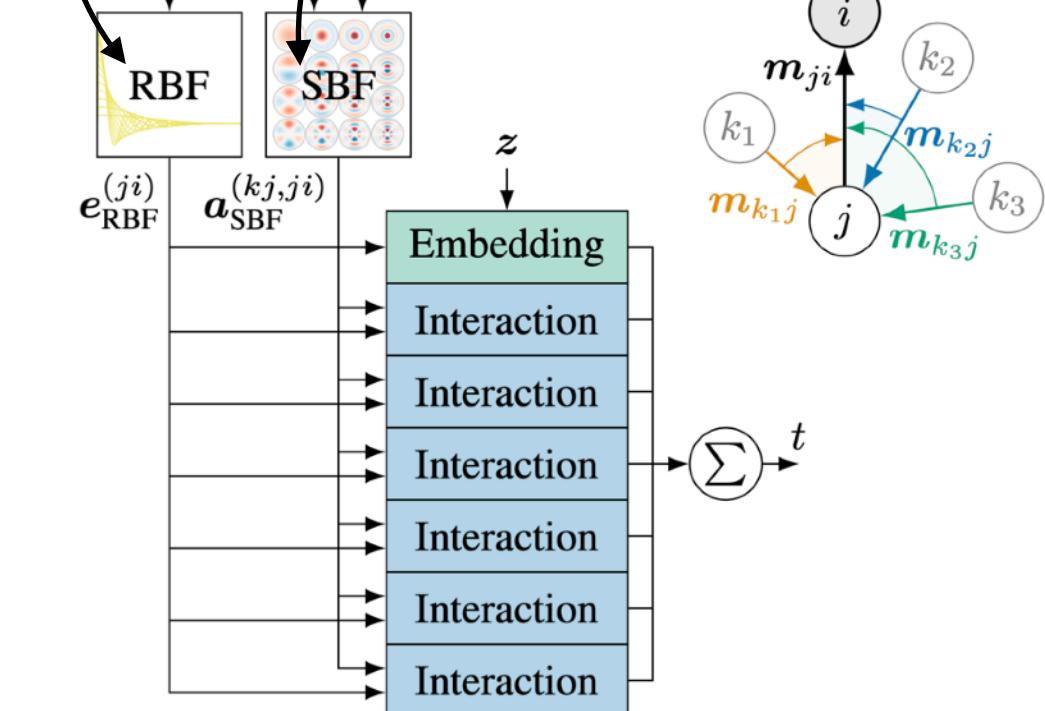
LightGBM w/ ECFP6
(without 3D geometry)



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



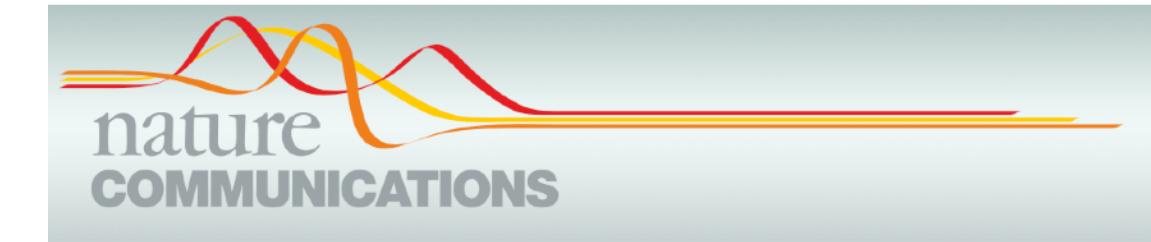
Radial Basis
辺上で更新
 $d_{ji} \downarrow$ $d_{kj} \downarrow$ $\alpha_{(kj,ji)} \downarrow$
角度も使う
 $\alpha_{(kj,ji)} = \angle x_k x_j x_i$



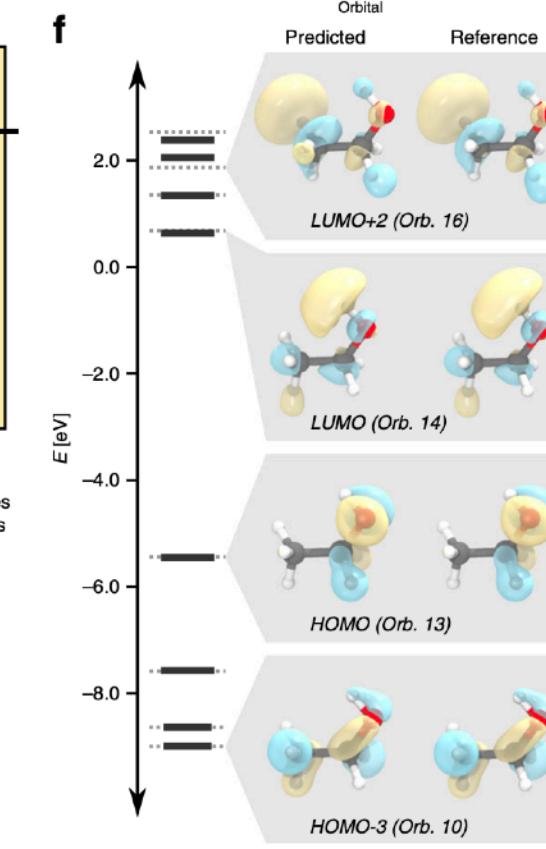
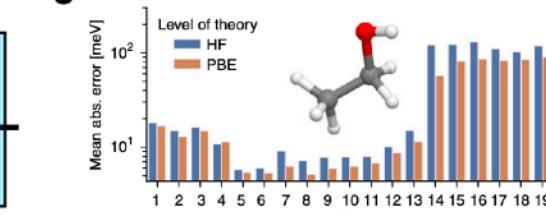
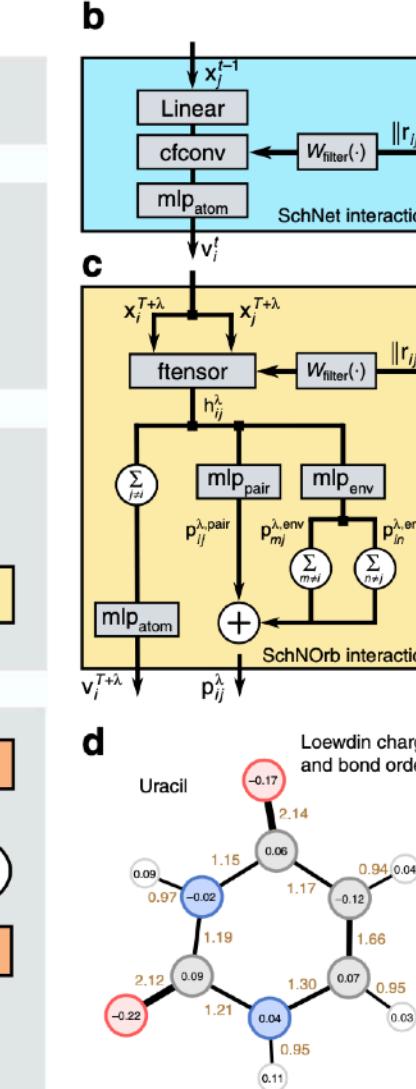
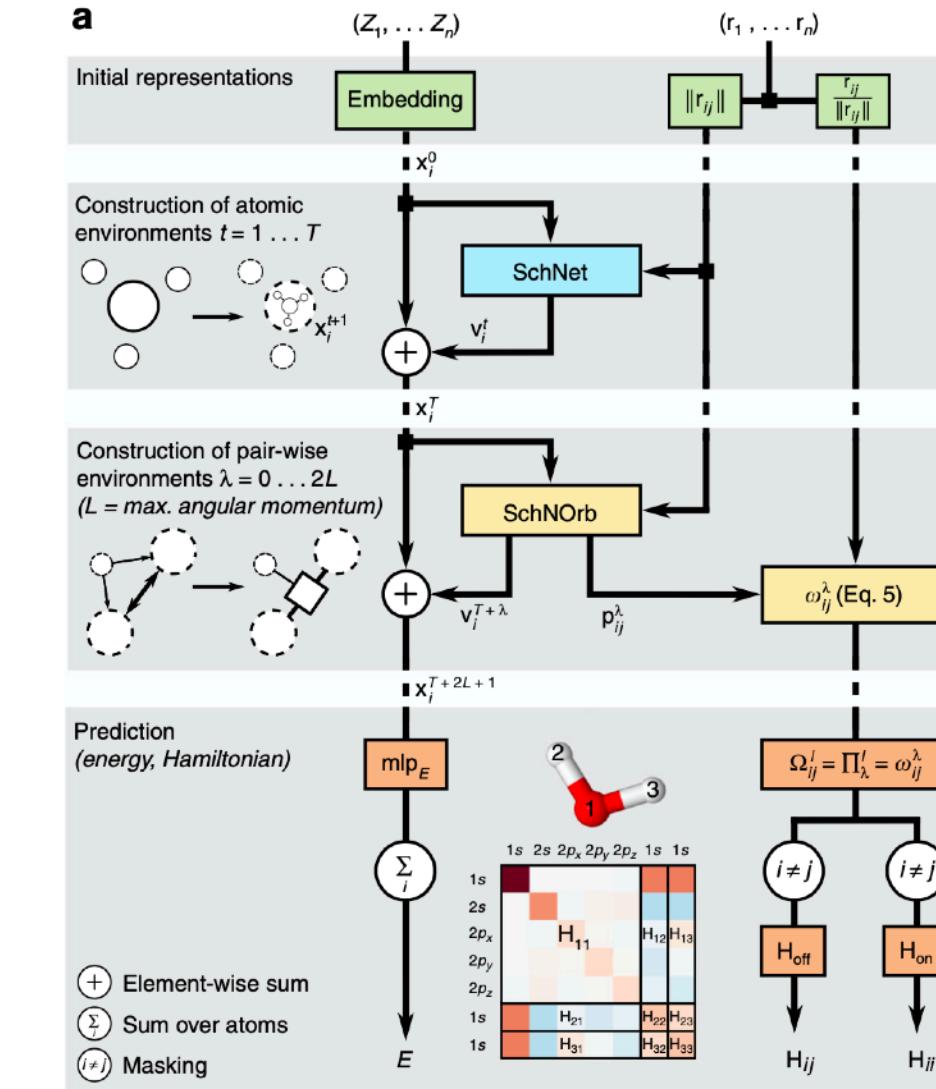
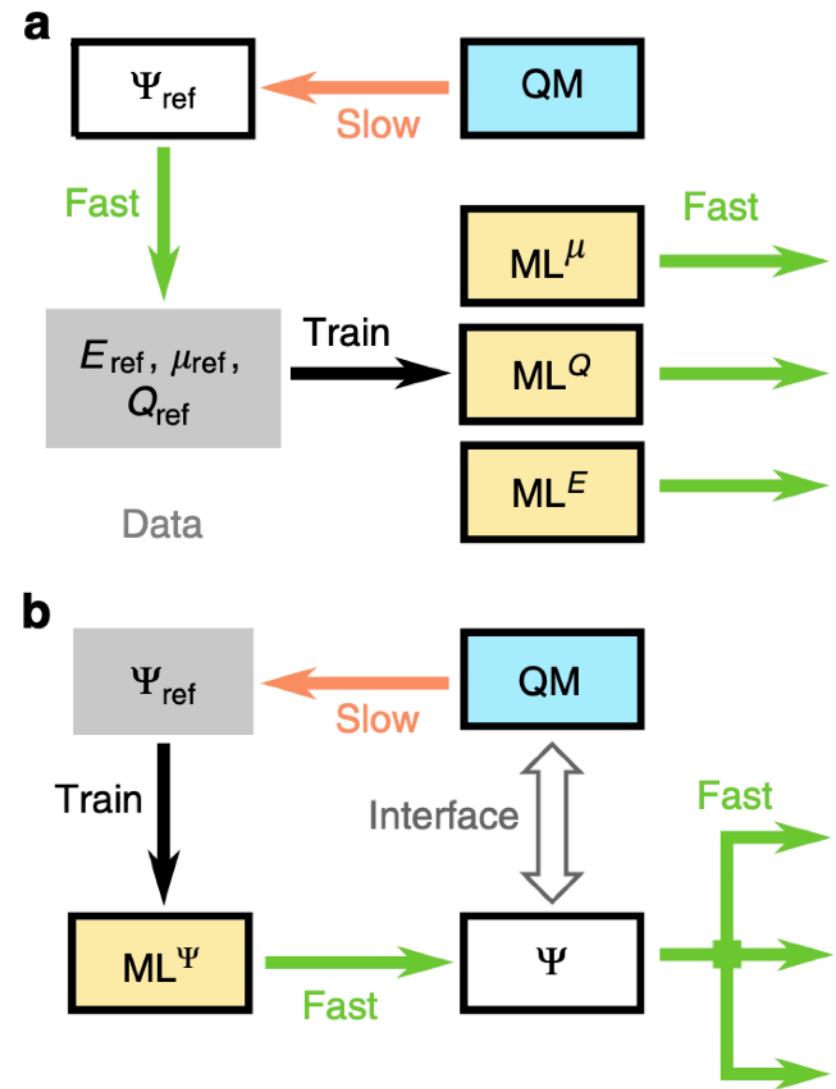
SchNOrb: 波動関数自体を機械学習

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

K. T. Schütt, M. Gastegger, A. Tkatchenko , K.-R. Müller & R. J. Maurer



Nature Communications 10, Article number: 5024 (2019)



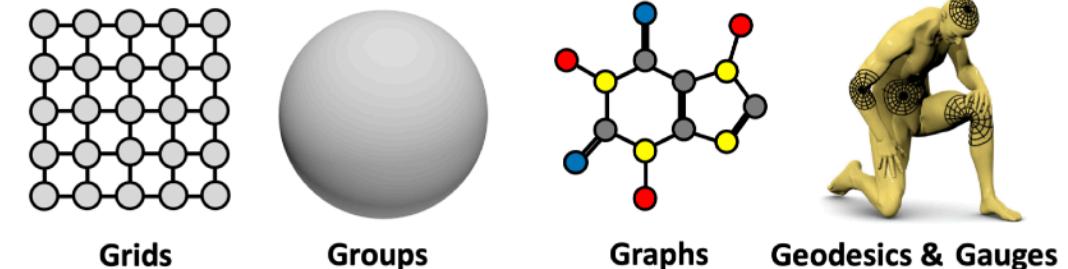
GNNから幾何的深層学習へ

[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

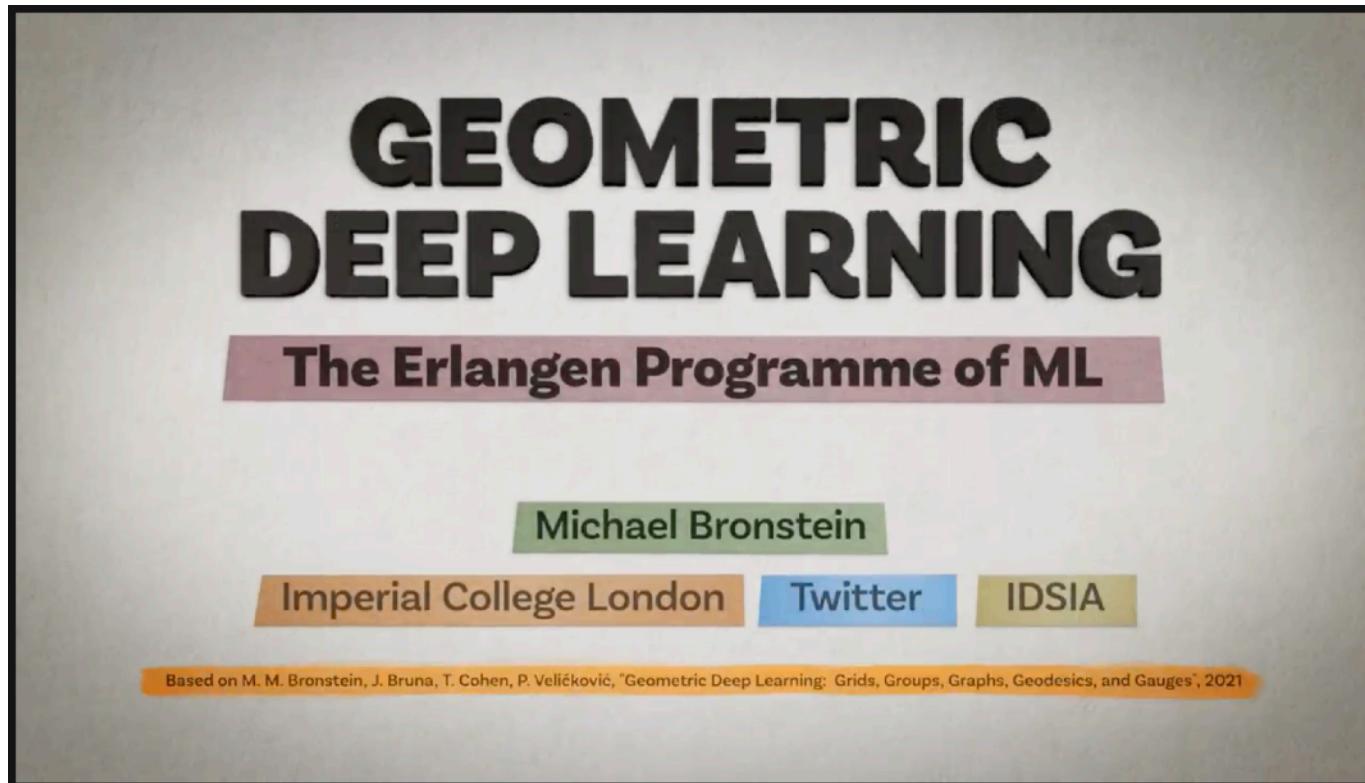
<https://arxiv.org/abs/2104.13478>



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!?)

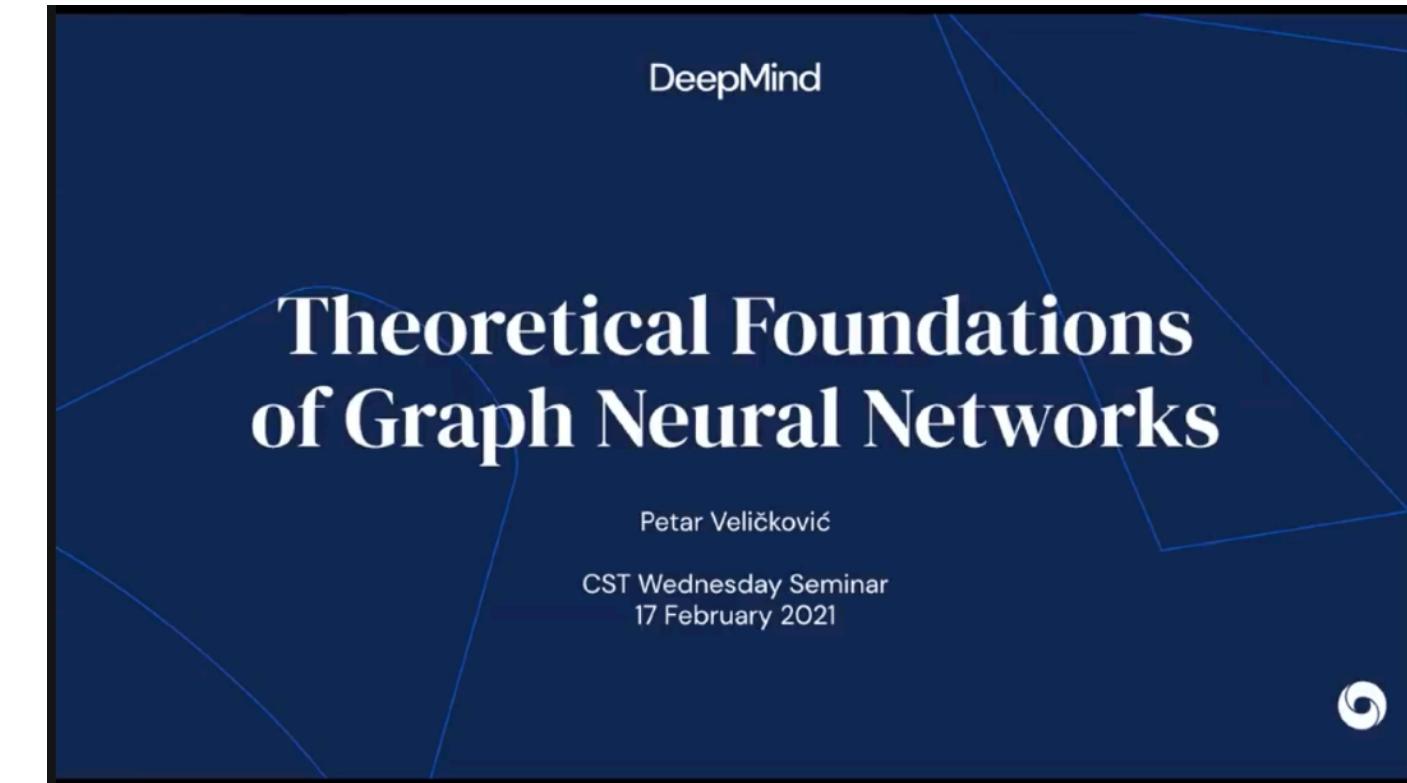
5Gs: Grids, Groups, Graphs, Geodesics/Gauges

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×

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

→結合長(原子間の距離)、結合角、二面角などを明示的に考慮する必要がある

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

頂点や辺の特徴量やGNN(=写像)のデザインで実現する

$E(3)$ 不変

- Schütt et al, [SchNet](#). (2017) <https://arxiv.org/abs/1706.08566>
- Unke et al, [PhysNet](#). (2019) <https://arxiv.org/abs/1902.08408>
- Klicpera et al, [DimeNet++](#). (2020) <https://arxiv.org/abs/2011.14115>

$SE(3)$ 同変

- Anderson et al, [Cormorant](#). (2019) <https://arxiv.org/abs/1906.04015>
- Fuchs et al, [SE\(3\)-Transformers](#). (2021) <https://arxiv.org/abs/2006.10503>

$E(3)$ 同変

- Thomas et al, [Tensor Field Networks](#). (2018) <https://arxiv.org/abs/1802.08219>
- Köhler et al, [Equivariant Flows \(Radial Field\)](#). (2020) <https://arxiv.org/abs/2006.02425>
- Satorras et al, [E\(n\) Equivariant Graph Neural Networks](#). (2021) <https://arxiv.org/abs/2102.09844>

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

不变 (invariant) $f(g \cdot x) = f(x)$

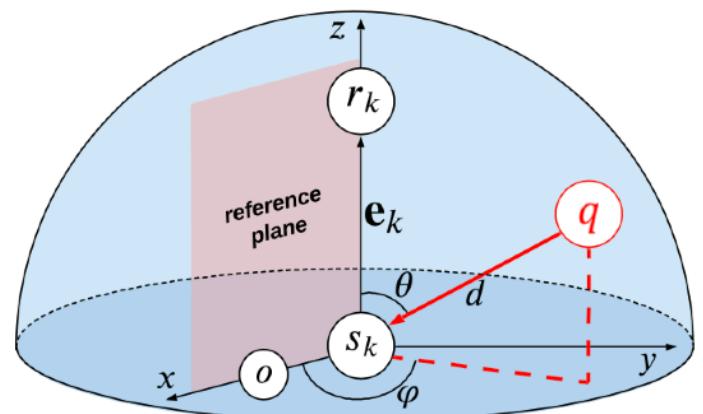
同変 (equivariant) $f(g \cdot x) = g \cdot f(x)$

研究は続く：さらに分子をより良く表現するには…

SphereNet

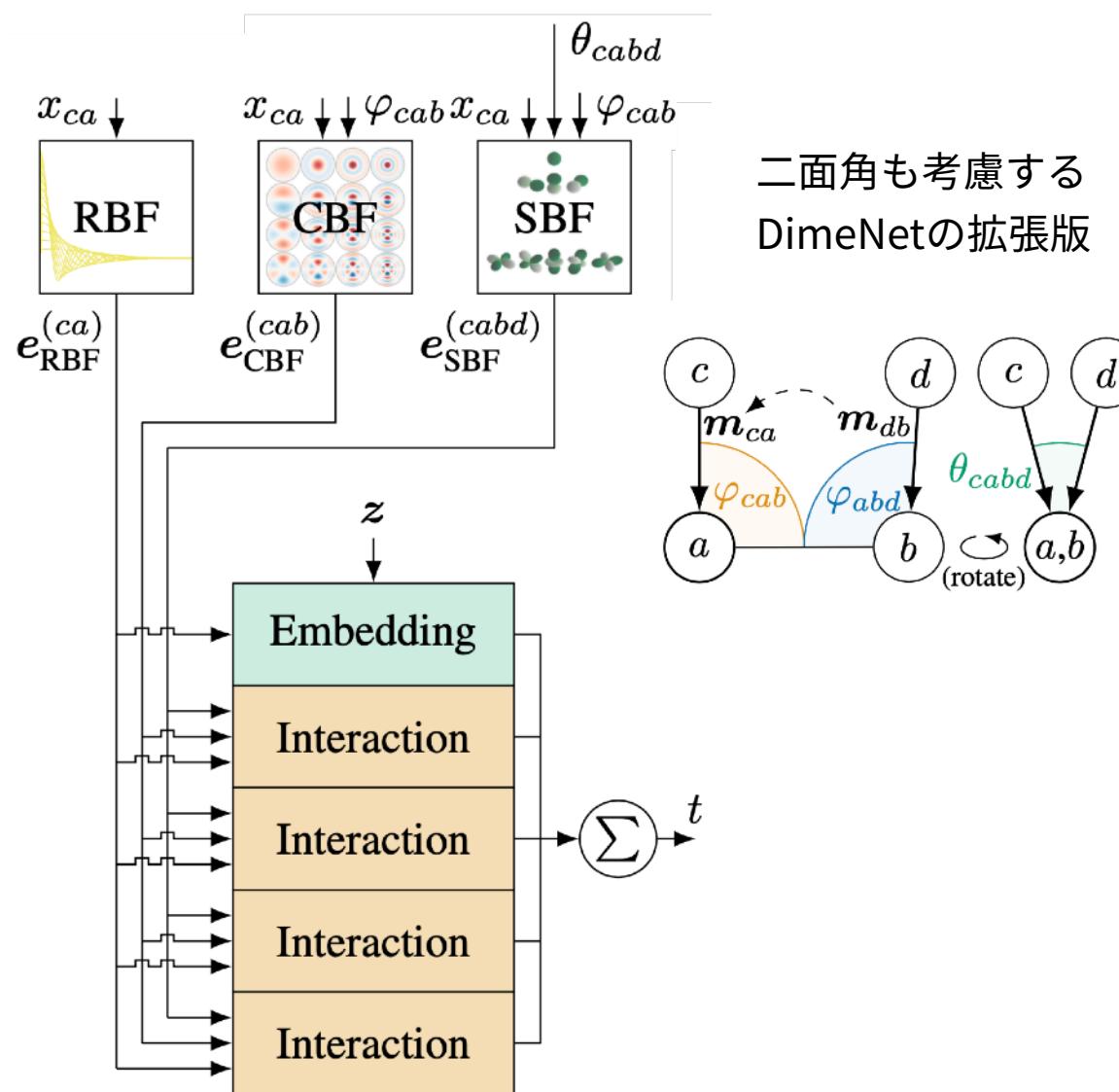
Liu et al
<https://arxiv.org/abs/2102.05013>

Spherical Message Passing (SMP)
として一般形を分析



GemNet

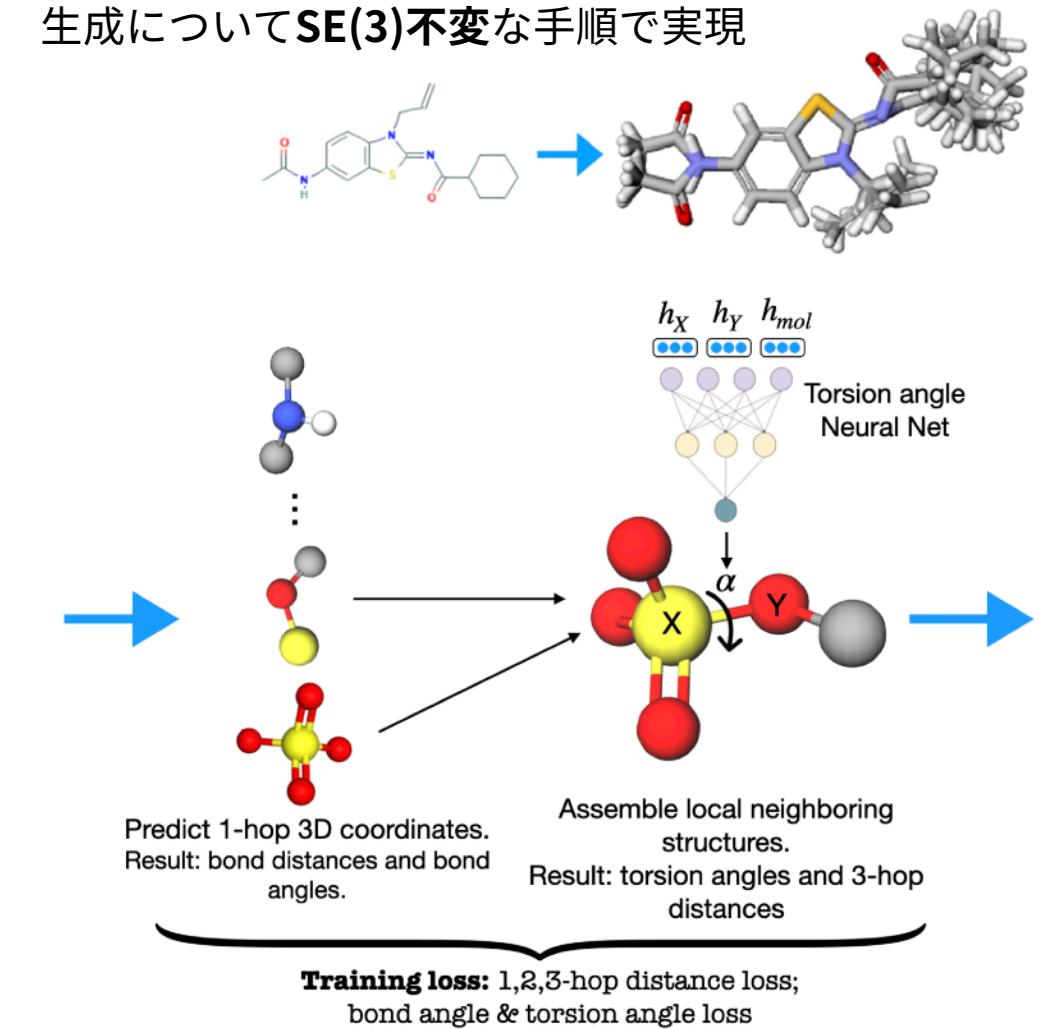
Klicpera et al (NeurIPS2021)
<https://arxiv.org/abs/2106.08903>



GeoMol

Ganea et al (NeurIPS2021)
<https://arxiv.org/abs/2106.07802>

低いエネルギーのConformerアンサンブルの
生成についてSE(3)不变な手順で実現



機械學習 × 量子化學計算

Machine Learning at the Atomic Scale (Chem. Rev.)
<https://pubs.acs.org/toc/chreay/121/16>

CHEMICAL REVIEWS

pubs.acs.org/CR

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



Review

Data Science Meets Chemistry (Acc. Chem. Res.)
<https://pubs.acs.org/page/achre4/data-science-meets-chemistry>

CHEMICAL REVIEWS

pubs.acs.org/CR

Physics-Inspired Structural Representations for Molecules and Materials

Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti*

 Cite This: *Chem. Rev.* 2021, 121, 9759–9815

 Read Online



Review

CHEMICAL REVIEWS

pubs.acs.org/CR

Ab Initio Machine Learning in Chemical Compound Space

Bing Huang and O. Anatole von Lilienfeld*

 Cite This: *Chem. Rev.* 2021, 121, 10001–10036

 Read Online



Review

ACCOUNTS of chemical research

pubs.acs.org/accounts

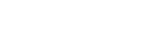
Article

Learning to Approximate Density Functionals

Published as part of the Accounts of Chemical Research special issue “*Data Science Meets Chemistry*”.
Bhupalee Kalita, Li Li, Ryan J. McCarty, and Kieron Burke*

 Cite This: *Acc. Chem. Res.* 2021, 54, 818–826

 Read Online



機械学習×シミュレーション

量子化学計算だけではなく様々な分野でシミュレーションと機械学習の融合研究が盛んに研究されるように

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



Annual Review of Physical Chemistry

Machine Learning for Molecular Simulation

Frank Noé,^{1,2,3} Alexandre Tkatchenko,⁴
Klaus-Robert Müller,^{5,6,7} and Cecilia Clementi^{1,3,8}

PNAS (2020)

The frontier of simulation-based inference

Kyle Cranmer^{a,b,1}, Johann Brehmer^{a,b}, and Gilles Louppe^c

^aCenter for Cosmology and Particle Physics, New York University, New York, NY 10003; ^bCenter for Data Science, New York University, New York, NY 10011;
and ^cMontefiore 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

Acc. Chem. Res. 54(7):1575–1585 (2021)



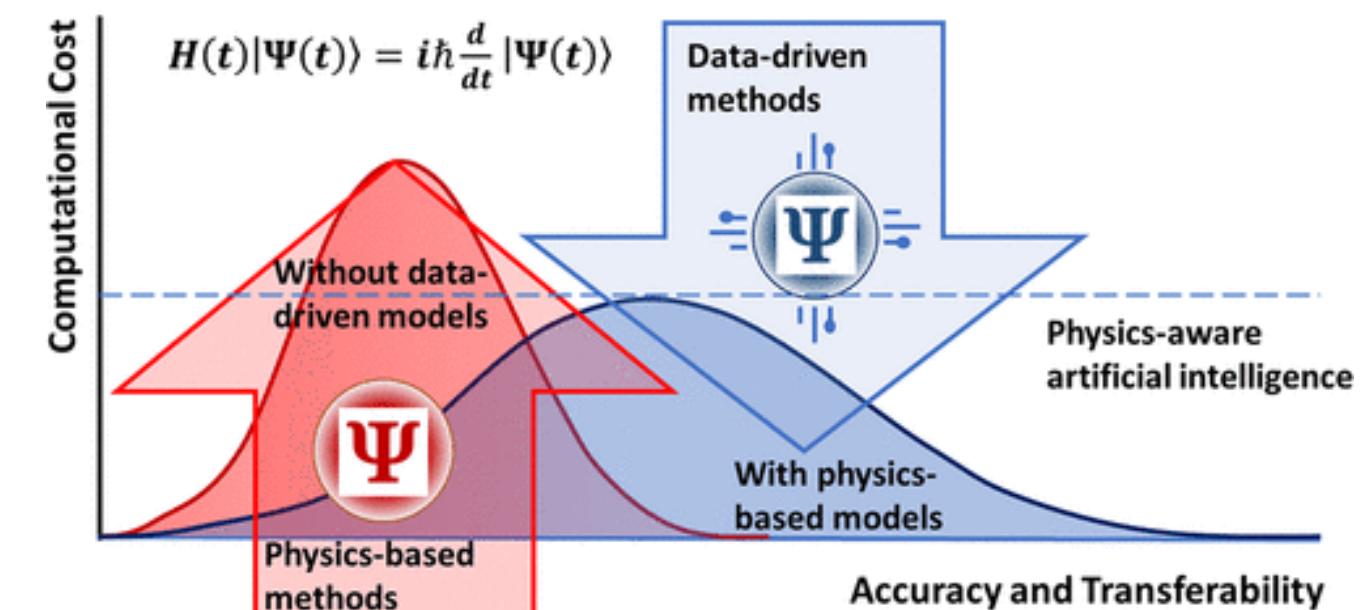
pubs.acs.org/accounts

Article

Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence

Published as part of the Accounts of Chemical Research special issue “Data Science Meets Chemistry”.

Tetiana Zubatiuk and Olexandr Isayev*



機械学習×手続き的・記号的操作/論理推論

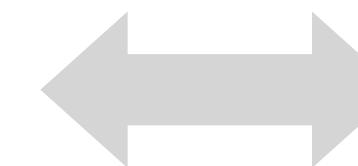
手続き的・記号的操作も学習できるプログラムとして扱えるようになってきた！

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)



明示的な化学知識も
融合していけるか？

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

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

Reaxys®

SCI-FINDER®
A CAS SOLUTION

CHEMATICÀ

ダークサイドへようこそ：こんにちは、世界！

✓ これまでの話は主に「量子化学計算によるデータ」でバーチャルな世界！

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・いろいろな大規模なオープンデータが利用できる (QM9, ANI1, OC20, PubChemQC, COLL, MD17, GuacaMol, ...)

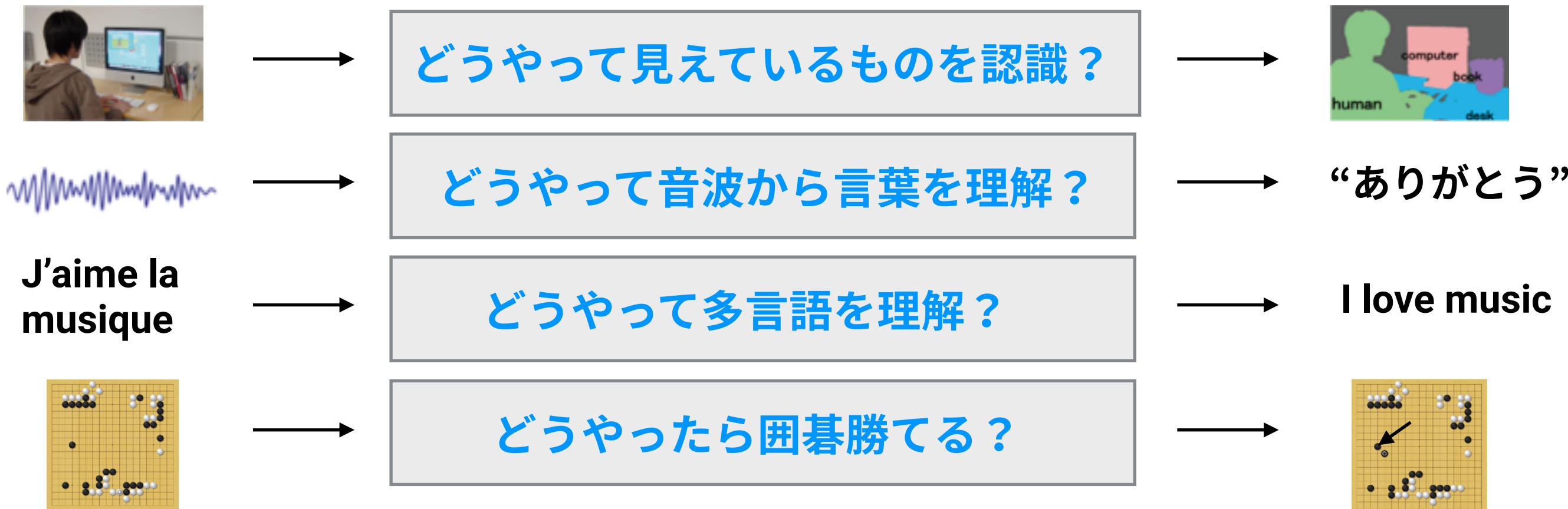
✓ リアルな世界はつらい… そんなに山ほどの同質なデータ取れないよ…

- 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ぬ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！！

機械学習×化学の真の問題

「予測ができる」ことは「理解」や「発見」ができるることを直接は意味しない！！

下記はどれも機械学習でかなり高精度な予測ができますが、それは私たちがその仕組みを理解できたことを少しでも意味するだろうか？



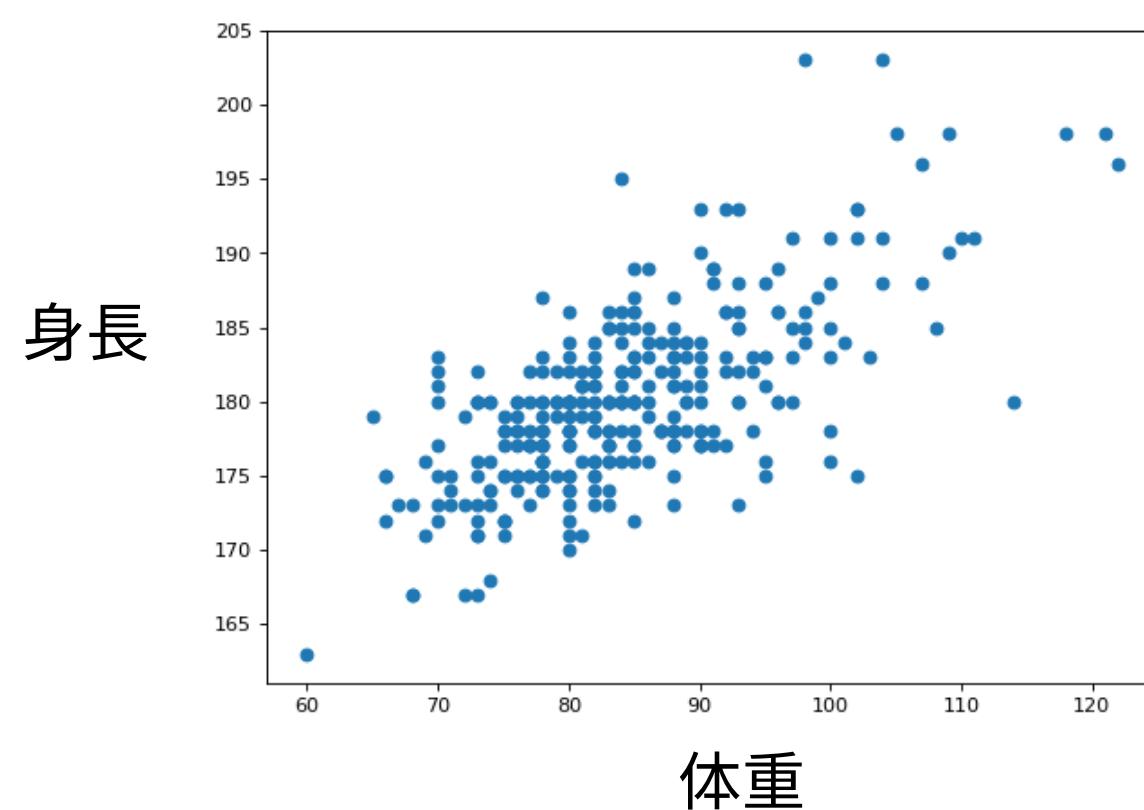
因果の理解には実験研究(介入研究)が必要不可欠

機械学習はあくまでデータの中の多次元相関を捉え、それによって予測する技術

→ 観察された相関が本当に因果性を含むのかを確かめるためには実験するしかない！

日本プロ野球開幕一軍選手の身長・体重データ

(2016年球団公式サイト選手データより自作)



「体重を増やせば身長も伸びる」が正しいかは
この観察データだけからは決して分からぬ

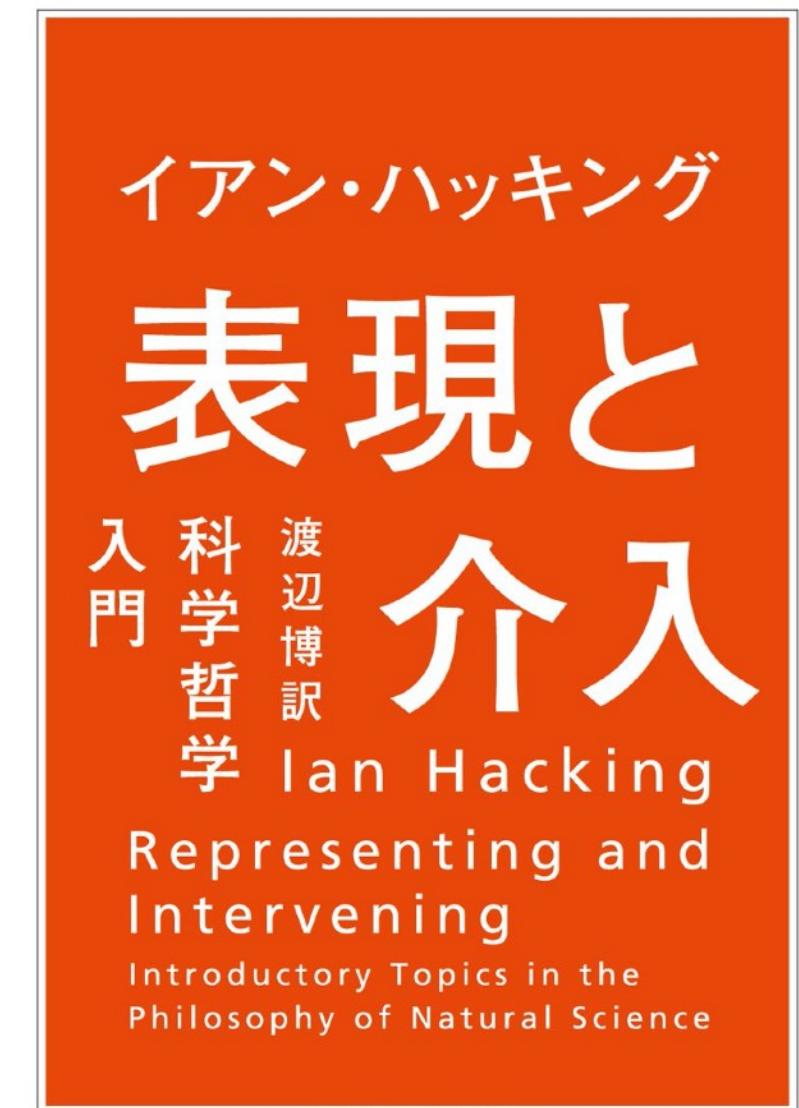
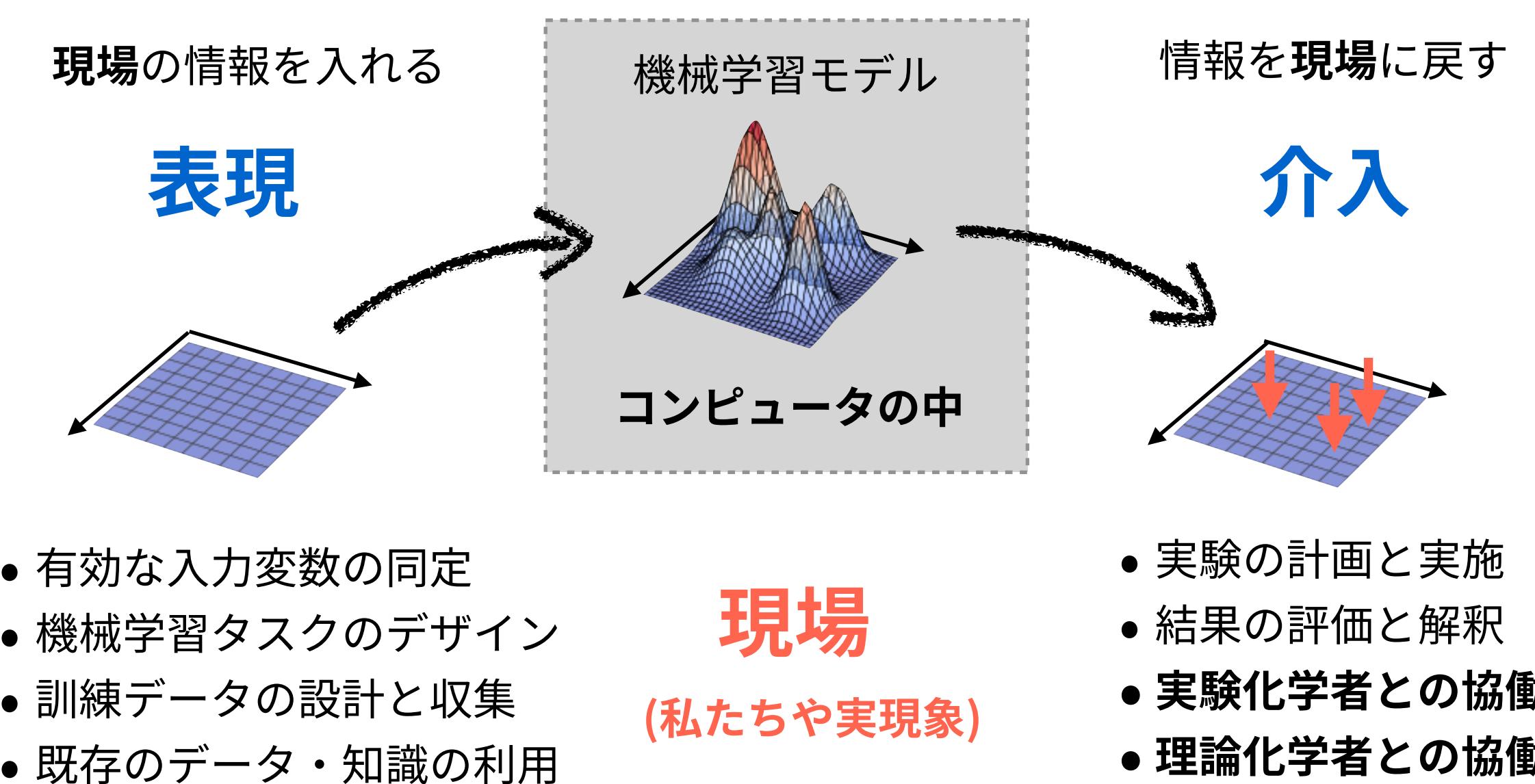
応用統計学の基本のキ

相関関係は必ずしも因果関係を意味しない

「予測ができる」ことは「理解」や「発見」ができる^{ことを直接は意味しない！！}

表現と介入：予測から理解・発見へ

事件はコンピュータ(機械学習)の中で起きてるんじゃない、**現場**で起きているんだ！ by 俺



ちくま学芸文庫

教訓：科学研究とは結局人間の営み！

「理解」や「発見」したいのは機械ではなく私たち人間

つまり、自然法則の問題ではなく**私たち自身の精神と世界のあり方の問題**を問うことになる！

教訓：科学研究とは結局人間の営み！

「理解」や「発見」したいのは機械ではなく私たち人間

つまり、自然法則の問題ではなく **私たち自身の精神と世界のあり方の問題** を問うことになる！

- 私たちの**ショボい認知能力**に収まるような「平易な理解」が求められている。

教訓：科学研究とは結局人間の営み！

「理解」や「発見」したいのは機械ではなく私たち人間

つまり、自然法則の問題ではなく **私たち自身の精神と世界のあり方の問題** を問うことになる！

- 私たちの**ショボい認知能力**に収まるような「平易な理解」が求められている。
- **有限の時間**しか生きられない私たちに「発見」という体験をお膳立てするためのヒント出しが求められている。（人類絶滅のタイムリミット内に）

教訓：科学研究とは結局人間の営み！

「理解」や「発見」したいのは機械ではなく私たち人間

つまり、自然法則の問題ではなく **私たち自身の精神と世界のあり方の問題** を問うことになる！

- 私たちの**ショボい認知能力**に収まるような「平易な理解」が求められている。
- **有限の時間**しか生きられない私たちに「発見」という体験をお膳立てするためのヒント出しが求められている。（人類絶滅のタイムリミット内に）
- **情報の部分性**：データにできる情報は**いつでも世界の情報量のほんのひとかけら**だけ。ゆく河の流れは絶えずして、しかももとの水にあらず。すべてを観測することはできない。

教訓：科学研究とは結局人間の営み！

「理解」や「発見」したいのは機械ではなく私たち人間

つまり、自然法則の問題ではなく **私たち自身の精神と世界のあり方の問題** を問うことになる！

- 私たちの**ショボい認知能力**に収まるような「平易な理解」が求められている。
- **有限の時間**しか生きられない私たちに「発見」という体験をお膳立てするためのヒント出しが求められている。（人類絶滅のタイムリミット内に）
- **情報の部分性**：データにできる情報は**いつでも世界の情報量のほんのひとかけら**だけ。ゆく河の流れは絶えずして、しかももとの水にあらず。すべてを観測することはできない。
- 人間が一生懸命集めたデータはどうしたって**何らかの偏り**から逃れられない。

教訓：科学研究とは結局人間の営み！

「理解」や「発見」したいのは機械ではなく私たち人間

つまり、自然法則の問題ではなく **私たち自身の精神と世界のあり方の問題** を問うことになる！

- 私たちの**ショボい認知能力**に収まるような「平易な理解」が求められている。
- **有限の時間**しか生きられない私たちに「発見」という体験をお膳立てするためのヒント出しが求められている。（人類絶滅のタイムリミット内に）
- **情報の部分性**：データにできる情報は**いつでも世界の情報量のほんのひとかけら**だけ。ゆく河の流れは絶えずして、しかももとの水にあらず。すべてを観測することはできない。
- 人間が一生懸命集めたデータはどうしたって**何らかの偏り**から逃れられない。

必要な情報のうち、**いつも偏った「一部」**しかデータにはできない前提で、私たち自身の許容限界に合う情報や示唆を得るために **「データを予測に変える道具」**をどう使えるか

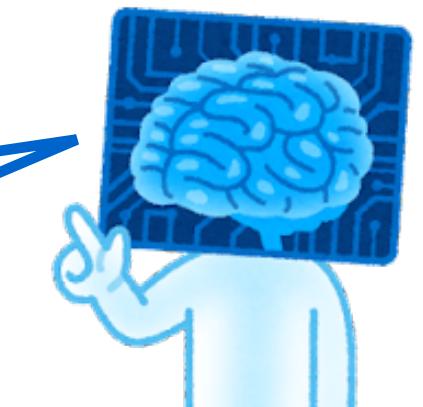
教訓：科学研究とは結局人間の営み！

「理解」や「発見」したいのは機械ではなく私たち人間

つまり、自然法則の問題ではなく**私たち自身の精神と世界のあり方の問題**を問うことになる！

- 私たちの**ショボい認知能力**に収まるような「平易な理解」が求められている。
- **有限の時間**しか生きられない私たちに「発見」という体験をお膳立てするためのヒント出しが求められている。（人類絶滅のタイムリミット内に）
- **情報の部分性**：データにできる情報は**いつでも世界の情報量のほんのひとかけら**だけ。ゆく河の流れは絶えずして、しかももとの水にあらず。すべてを観測することはできない。
- 人間が一生懸命集めたデータはどうしたって**何らかの偏り**から逃れられない。

ショボい認知能力のおまえら人間にとったら「ビッグ」データかもしらんけど、ホンマに必要な情報量からしたらハナクソみたいなもんやな！



機械学習から機械発見へ

実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



Science 363 (2019)



Nature 583 (2020)



Nature Reviews Drug Discovery 17 (2018)



機械学習から機械発見へ

実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



Science 363 (2019)



Nature 583 (2020)



Nature Reviews Drug Discovery 17 (2018)



- **機械発見技術の研究基盤として非常に重要**：再現性・属人性などデータの質と量の確保
+ 失敗データを取る実験やランダム実験はデータ科学上は必要だが人間はやりたくない…

機械学習から機械発見へ

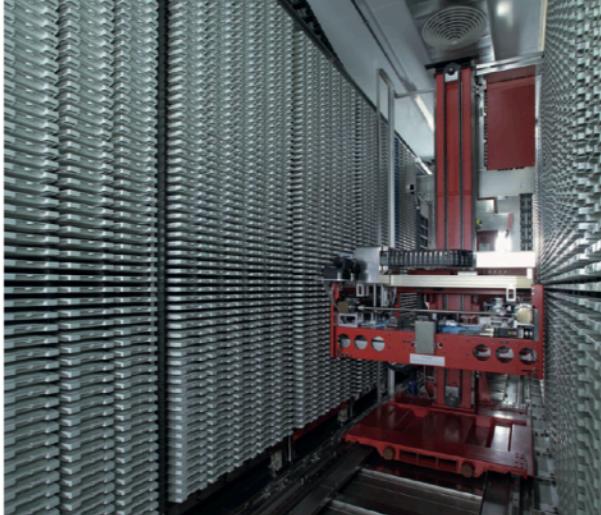
実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



Science 363 (2019)



Nature 583 (2020)



Nature Reviews Drug Discovery 17 (2018)



- **機械発見技術の研究基盤として非常に重要**：再現性・属人性などデータの質と量の確保
+ 失敗データを取る実験やランダム実験はデータ科学上は必要だが人間はやりたくない…
- **発見が自動化できるか**はAI分野にとっても積年の未解決問題。「人工知能」を作りたいなら私たちが日々小さな「発見」と「学習」を繰り返して世界を理解していく過程の理解は不可避

機械学習から機械発見へ

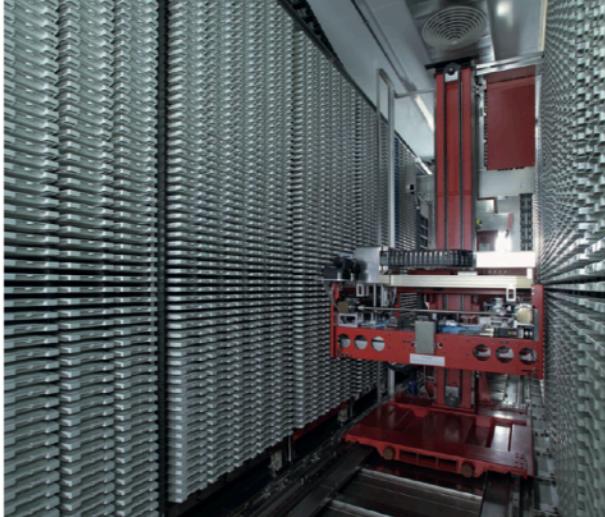
実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



Science 363 (2019)



Nature 583 (2020)



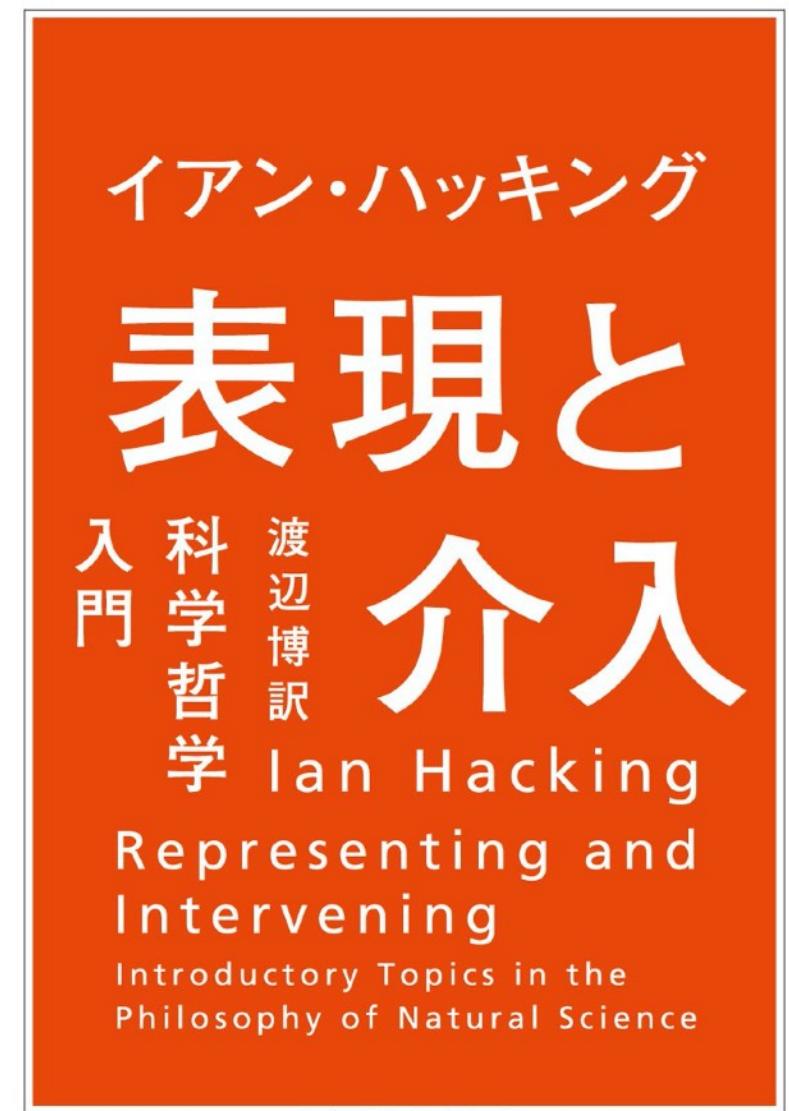
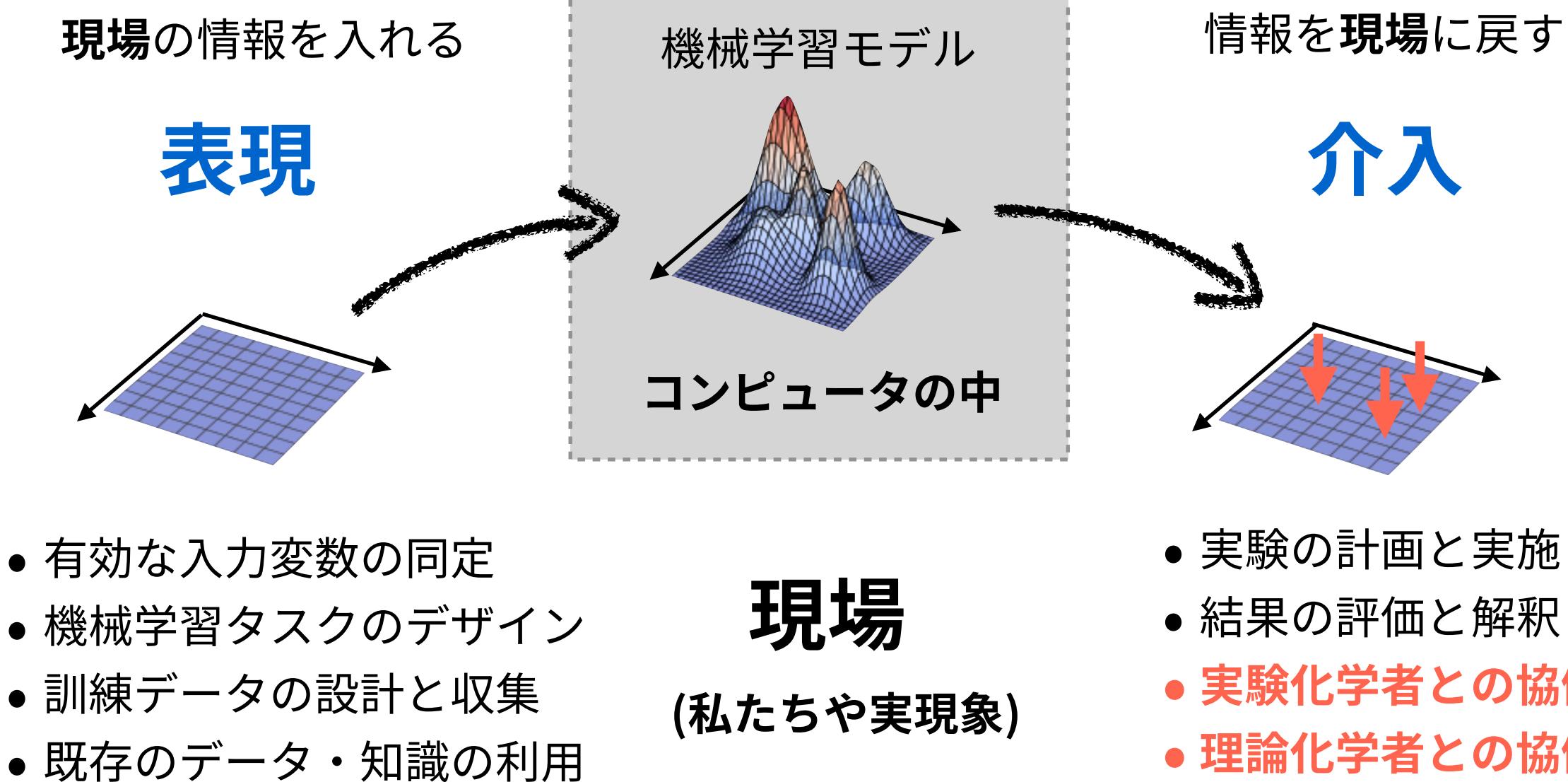
Nature Reviews Drug Discovery 17 (2018)



- **機械発見技術の研究基盤として非常に重要**：再現性・属人性などデータの質と量の確保
+ 失敗データを取る実験やランダム実験はデータ科学上は必要だが人間はやりたくない…
- **発見が自動化できるか**はAI分野にとっても積年の未解決問題。「人工知能」を作りたいなら私たちが日々小さな「発見」と「学習」を繰り返して世界を理解していく過程の理解は不可避
- 実験自動化が実現されても 「常にひとかけらの部分情報しか手に入らない」 本質は**変わらない**

機械学習屋よ、「頭でっかち」から脱し、現場に出よう！

事件はコンピュータ(機械学習)の中で起きてるんじゃない、現場で起きているんだ！ by 俺



北海道大学 化学反応創成研究拠点(ICReDD)^{アイクレッド}

百戦錬磨の計算化学者・実験化学者・情報科学者が結託して「**化学反応のデザインと発見**」のやり方を革新することを目指して集う梁山泊。日々楽しい研究と議論が繰り広げられている。

私のような技術屋にとっても**機械学習・機械発見**の技術研究と実世界検証のための胸アツな**現場**



まとめ：May the ML force be with you...

機械学習と機械発見の技術屋から見たデータ中心型の化学・材料科学の教訓とこれから

ライトサイド（光明面）

機械学習は「データを予測に変える」強力なテクノロジー！

- 分子の表現学習とGraph Neural Networks
- 帰納バイアスの設計とグレイボックス最適化

ダークサイド（暗黒面）

自然科学の実現象データで使うのはいろいろ激ムズ！！！

- 羅生門効果とUnderspecification
- 「予測ができる」とは「理解」や「発見」ができる意味しない！

人が事実を用いて科学をつくるのは、石を用いて家を造るようなものである。
事実の集積が科学でないことは、石の集積が家でないのと同じことである。

アンリ・ポアンカレ「科学と仮説」

