

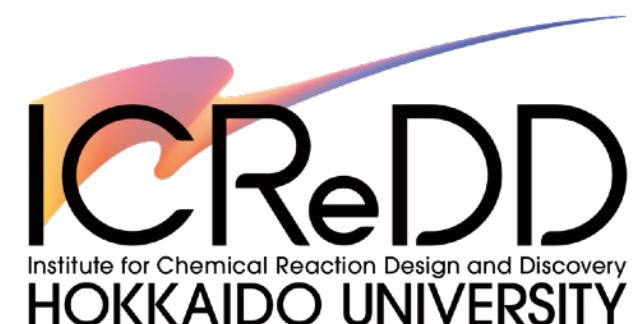
機械学習・機械発見から見る データ中心型化学の野望と憂鬱

2021年12月9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@京阪奈ATR
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！

- うどん県高松市生まれ
- 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)

普段は京大iPS細胞研との連携ラボ@京阪奈ATRに勤務
(iPS細胞連携医学的リスク回避チーム)



たきがわ いちがく
瀧川 一学

<https://itakigawa.github.io>

と同時に機械学習のユーザでもあります！

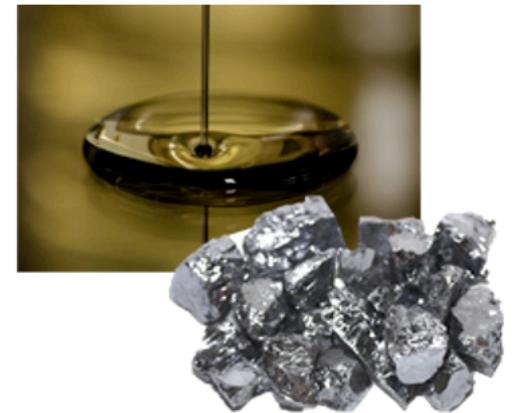
- うどん県高松市生まれ
 - 1995～2004 北海道大 (工学研究科)
2004 博士(工学) "劣決定信号源分離の解の理論分析"
 - 2005～2011 京都大 (化学研究所/薬学研究科)
バイオインフォマティクスセンター 助教
 - 2012～2018 北海道大 (情報科学研究科)
大規模知識処理研究室 准教授
2015～2018 JSTさきがけ (材料インフォマティクス)
 - 2019～ 北海道大学 化学反応創成研究拠点(ICReDD)
2019～ 理化学研究所 革新知能統合研究センター(AIP)
- 普段は京大iPS細胞研との連携ラボ@京阪奈ATRに勤務
(iPS細胞連携医学的リスク回避チーム)

最小限の前おき：化学=物質AをBに変えたい！

原子がひとまとめになった「分子」とその化学反応は広範囲な分野の主人公

化学、医学・生理学、物理学、生物学、創薬、材料科学、環境科学、農学、食品、化粧品、…

身の回りのもの



私たち自身(生命現象)

Chemical
Reactions



"化"学

エネルギー



分子は「組合せ的」側面を持つ

c&en

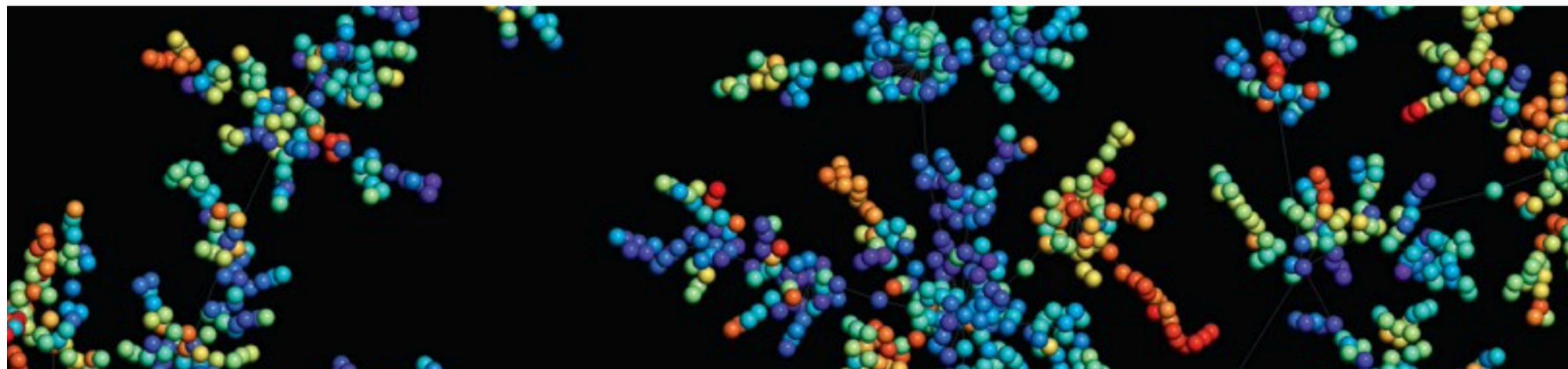
COMPUTATIONAL CHEMISTRY

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}

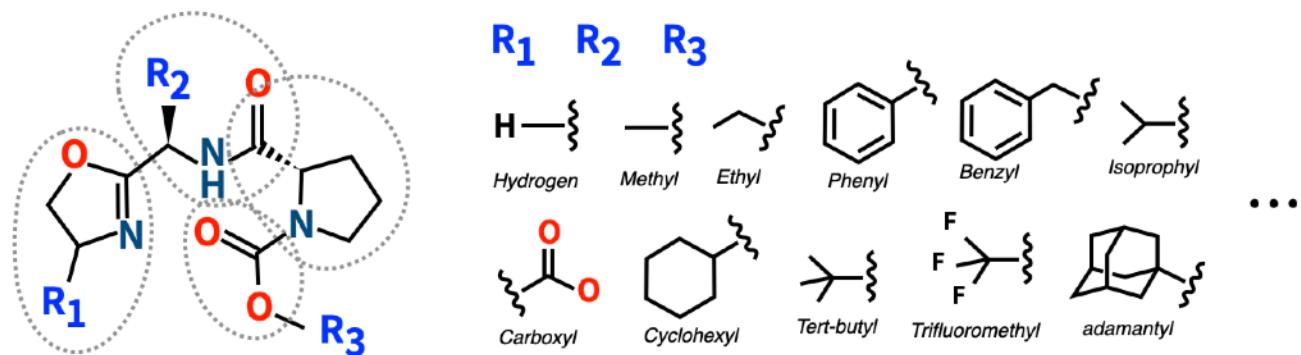
An estimate of the number of possible small organic molecules

10^8

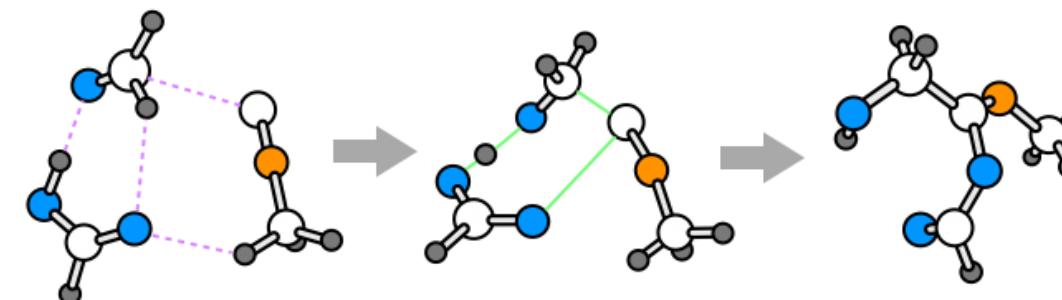
The number of organic and inorganic substances in the CAS database

機械学習で化学をやってみた！

分子は原子間の結合の組合せ（組合せ規則は量子力学）→技術上の関心は「組合せの機械学習」



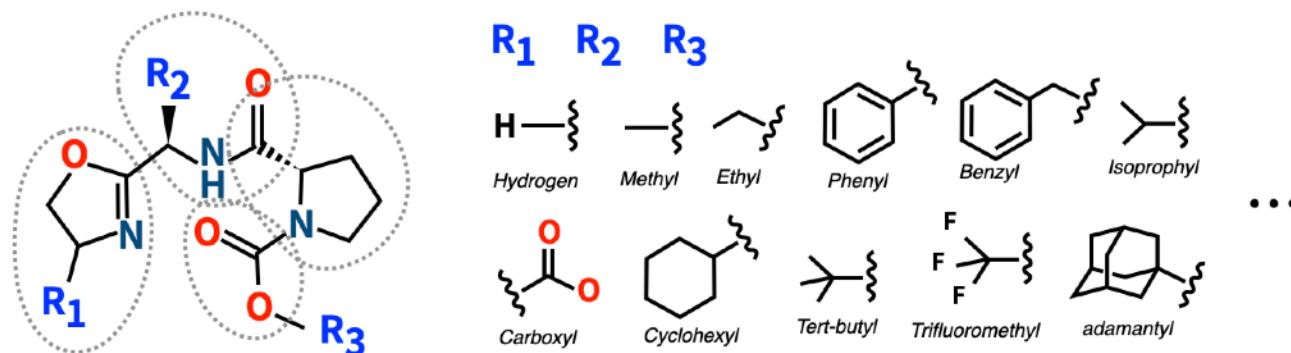
要素の組合せに構成性と階層性がある（言語に似ている？）



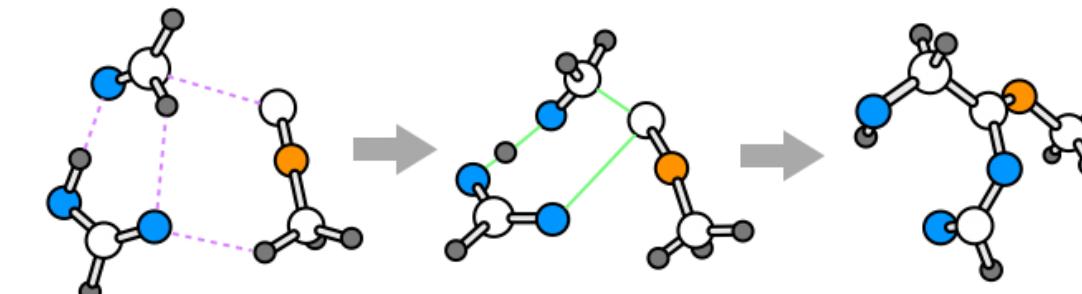
化学反応は結合の組み替え（文法に似ている？）

機械学習で化学をやってみた！

分子は原子間の結合の組合せ（組合せ規則は量子力学）→ 技術上の関心は「組合せの機械学習」



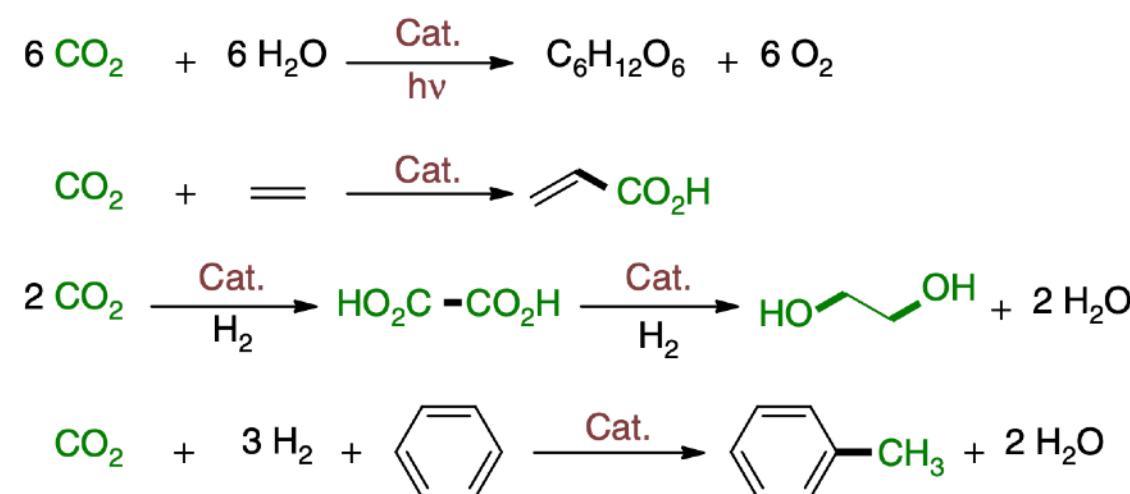
要素の組合せに構成性と階層性がある（言語に似ている？）



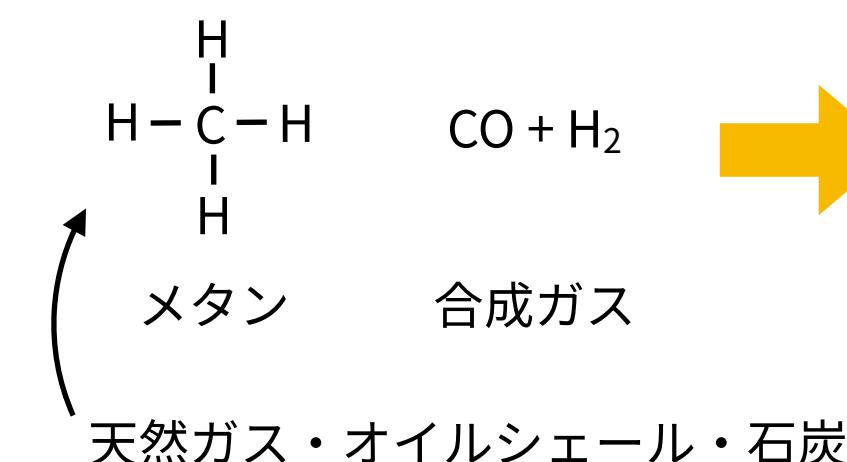
化学反応は結合の組み替え（文法に似ている？）

本当のゴールは「化学反応のデザインと発見」で機械学習はそのための道具の一つ

CO₂の資源利用：人工光合成



C1化学 原料はCが一つ



C2炭化水素類の合成

エチレン: C ₂ H ₄	エタン: C ₂ H ₆
$\begin{array}{c} \text{H} \\ \\ \text{H}-\text{C}=\text{C}-\text{H} \\ \\ \text{H} \end{array}$	$\begin{array}{c} \text{H} \text{ H} \\ \quad \\ \text{H}-\text{C}-\text{C}-\text{H} \\ \quad \\ \text{H} \text{ H} \end{array}$

天然ガス・オイルシェール・石炭・バイオマスなど石油以外から得られる

今日の話：機械学習の力の光明面と暗黒面

Q. 機械学習・機械発見の技術って**自然現象の理解・発見**の役に立つのかな？(化学を例に)

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

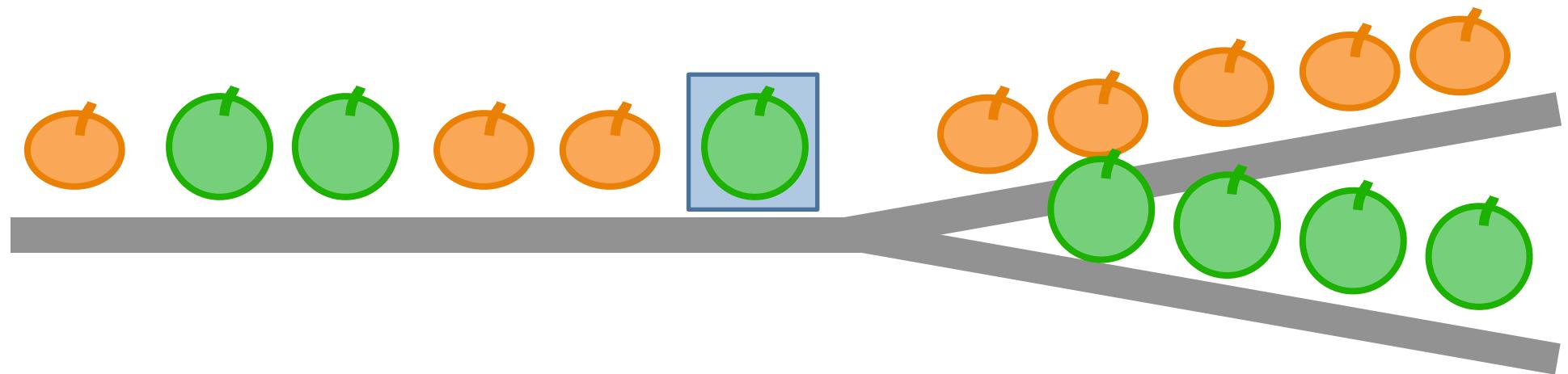
ダークサイド（暗黒面）

自然科学の実現象データで使うのはいろいろ激ムズ！！！

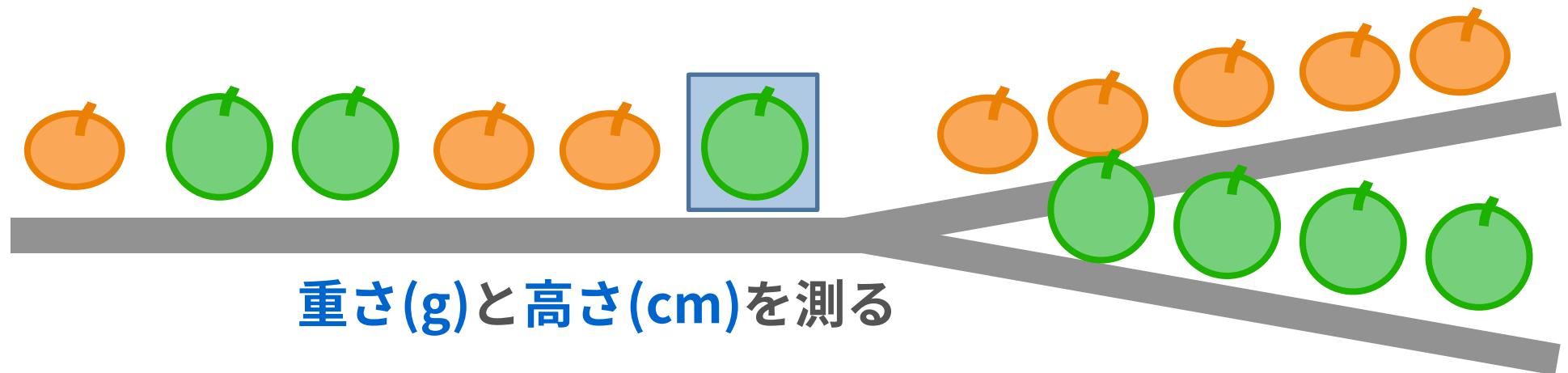
- ✓ 羅生門効果とUnderspecification
- ✓ 「予測ができること」は「理解」や「発見」ができるることを意味しない！

フォースと共にあらんことを *May the ML force be with you...*

機械学習は「データを予測に変える」技術

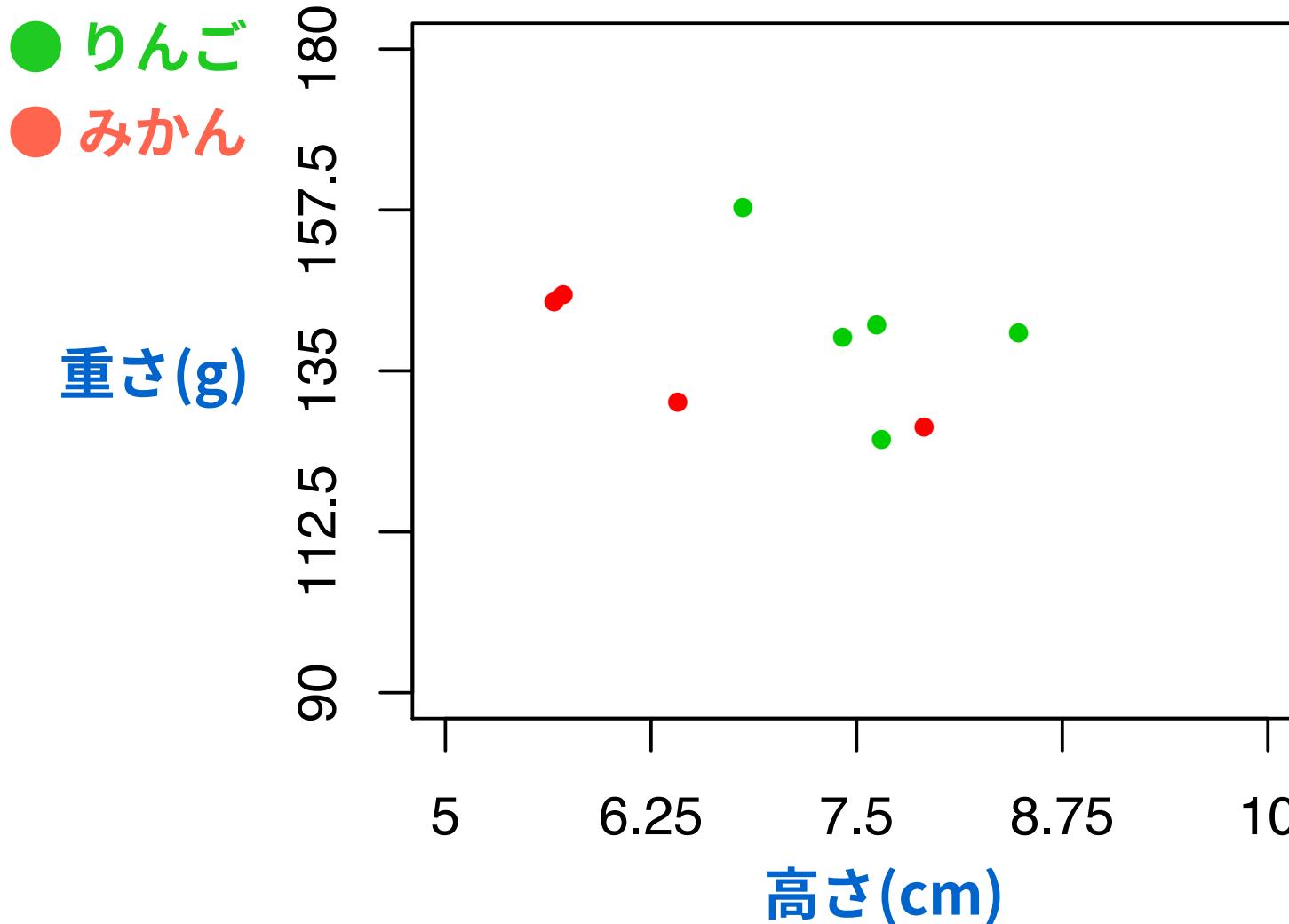
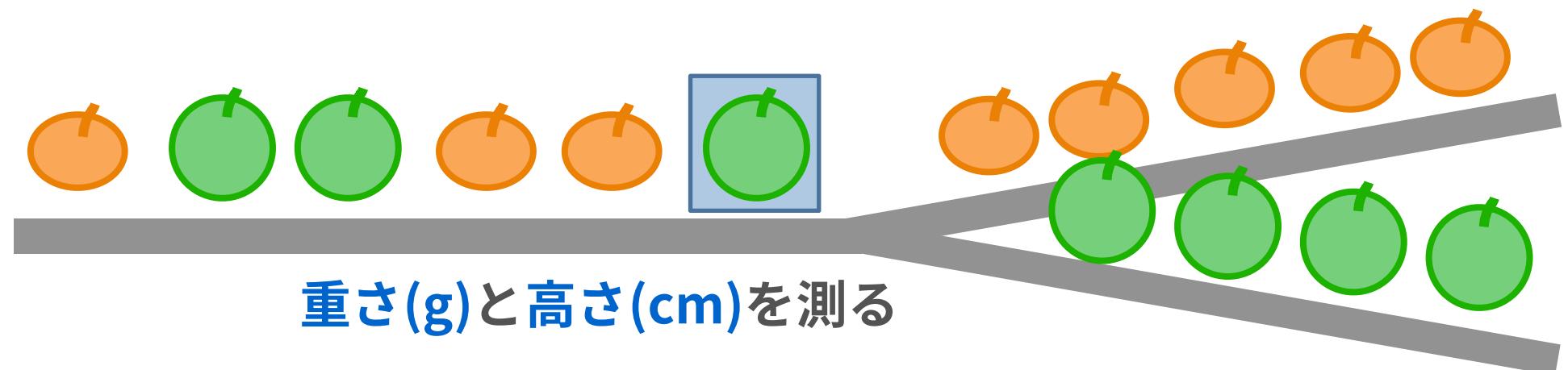


機械学習は「データを予測に変える」技術

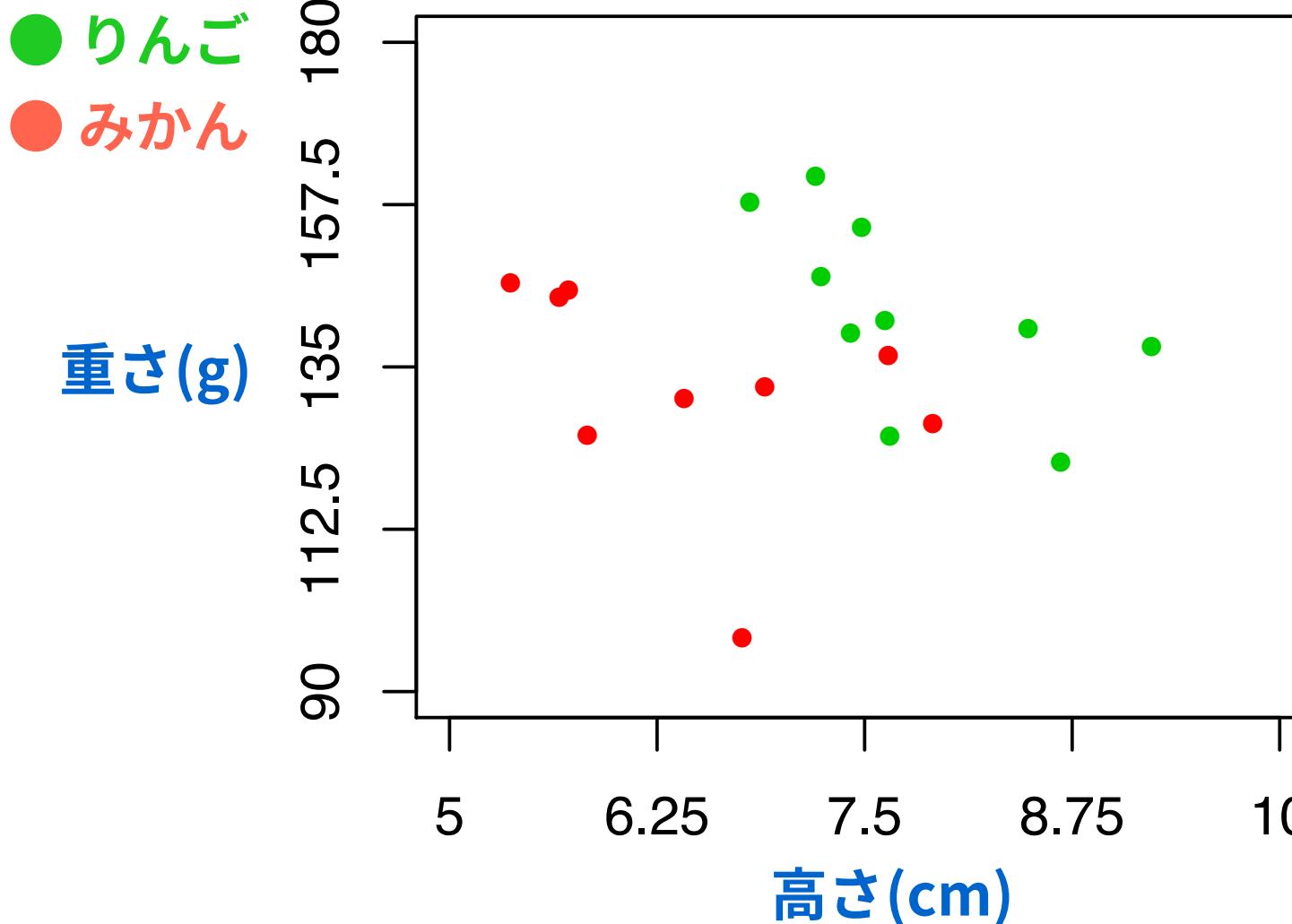
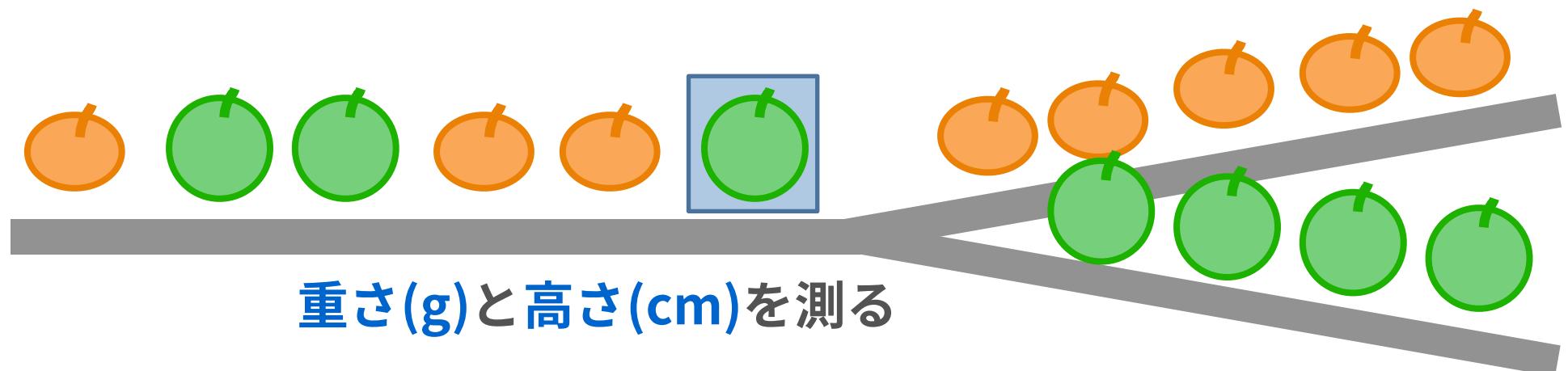


重さ(g)と高さ(cm)を測る

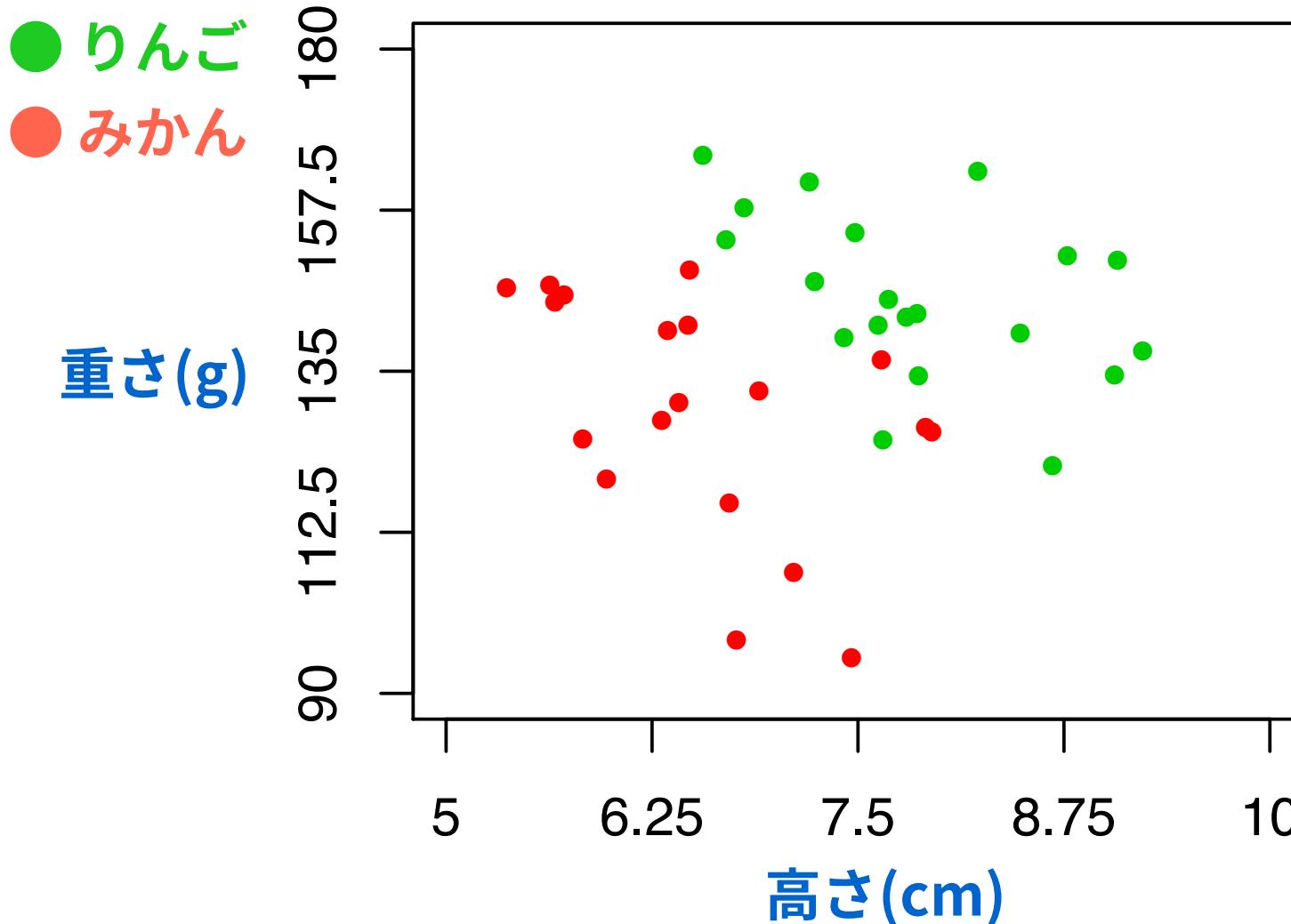
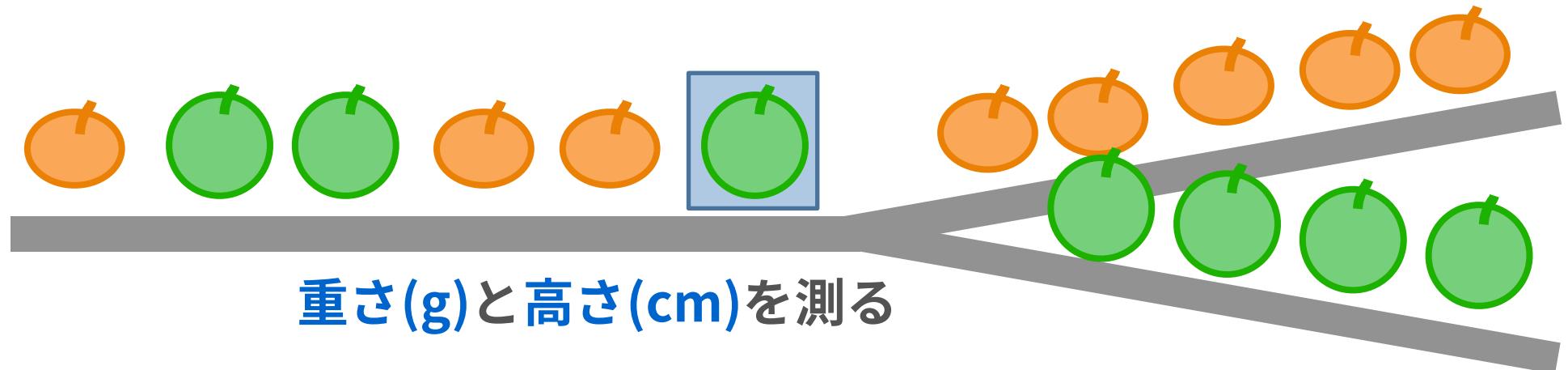
機械学習は「データを予測に変える」技術



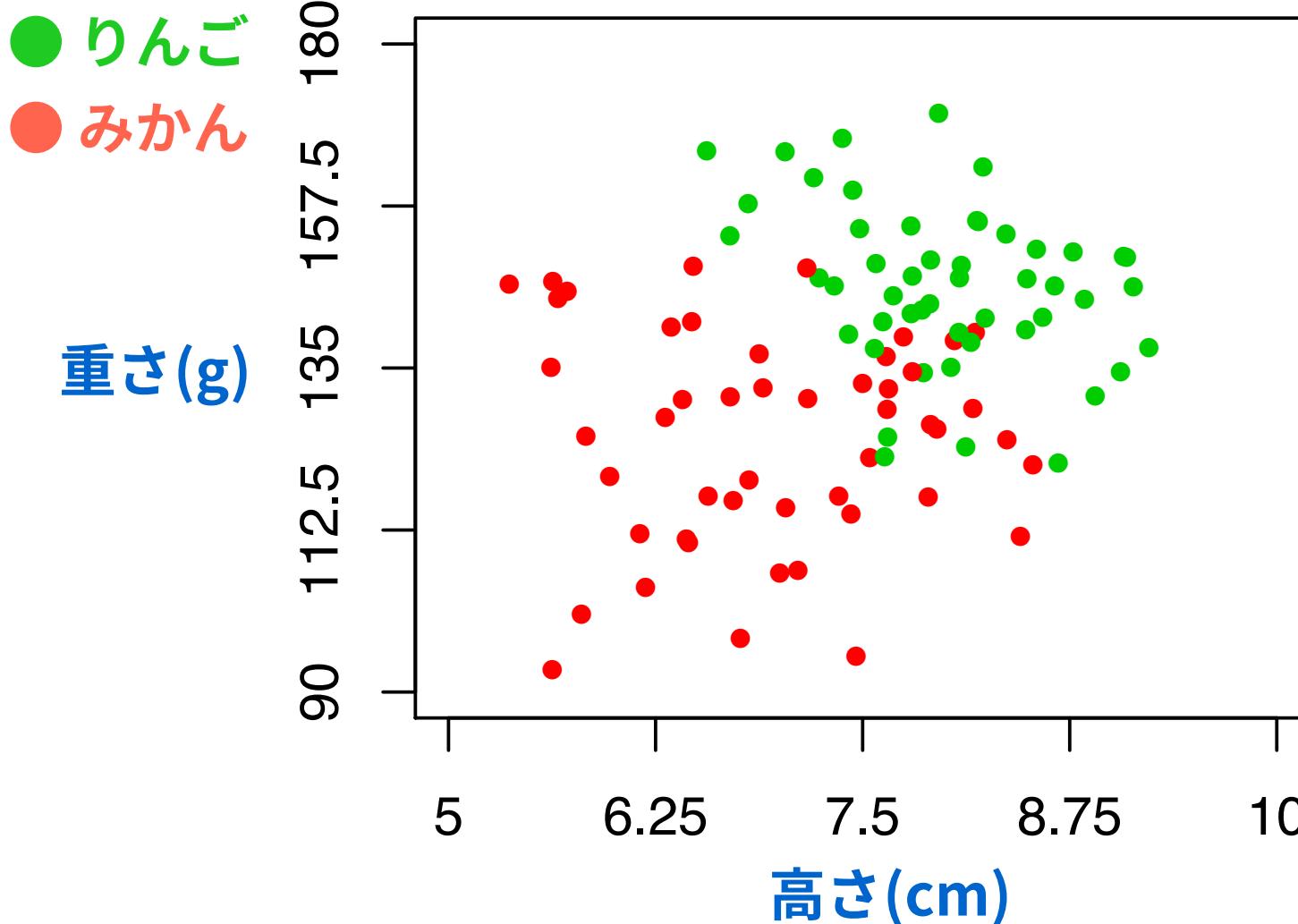
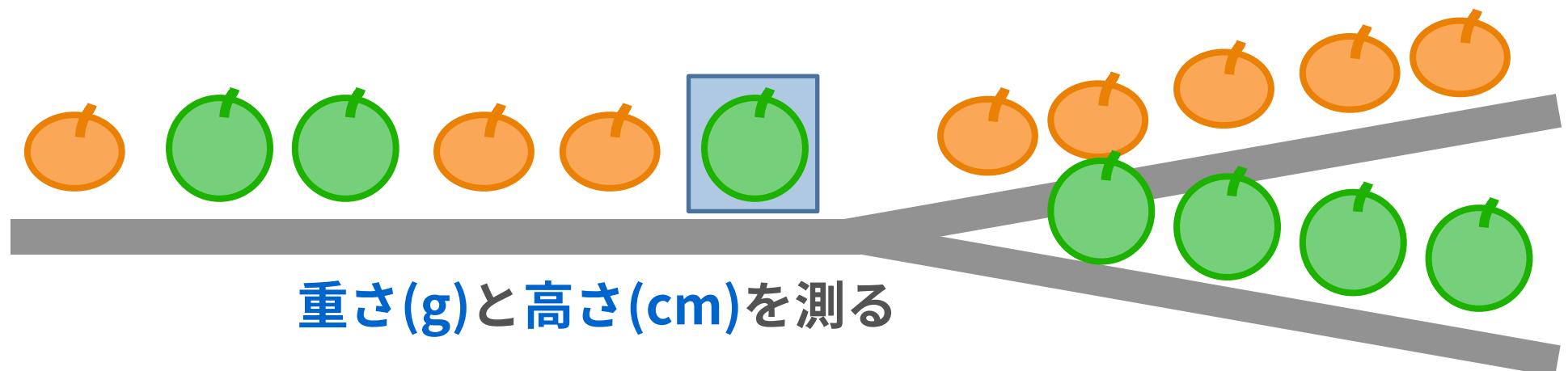
機械学習は「データを予測に変える」技術



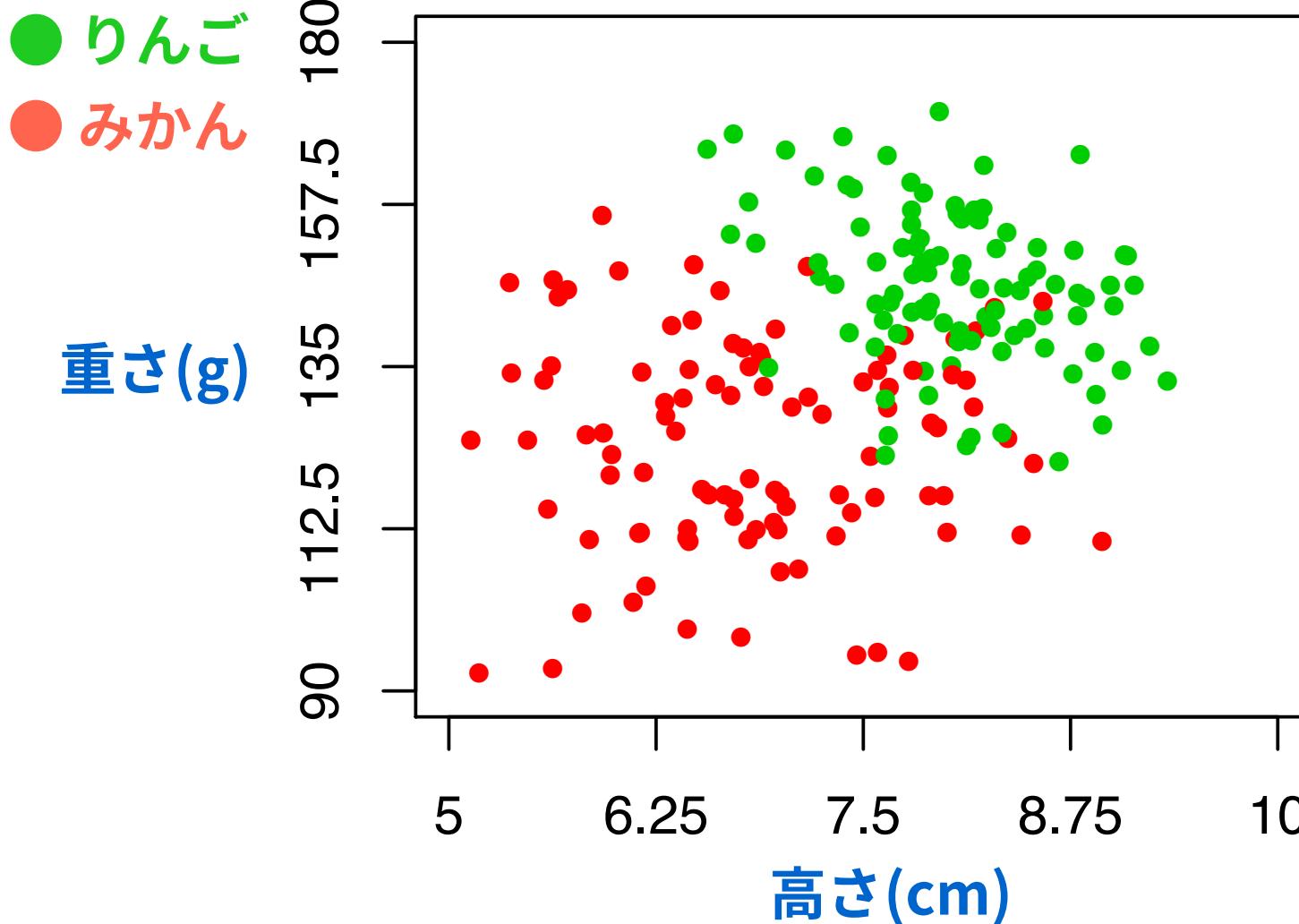
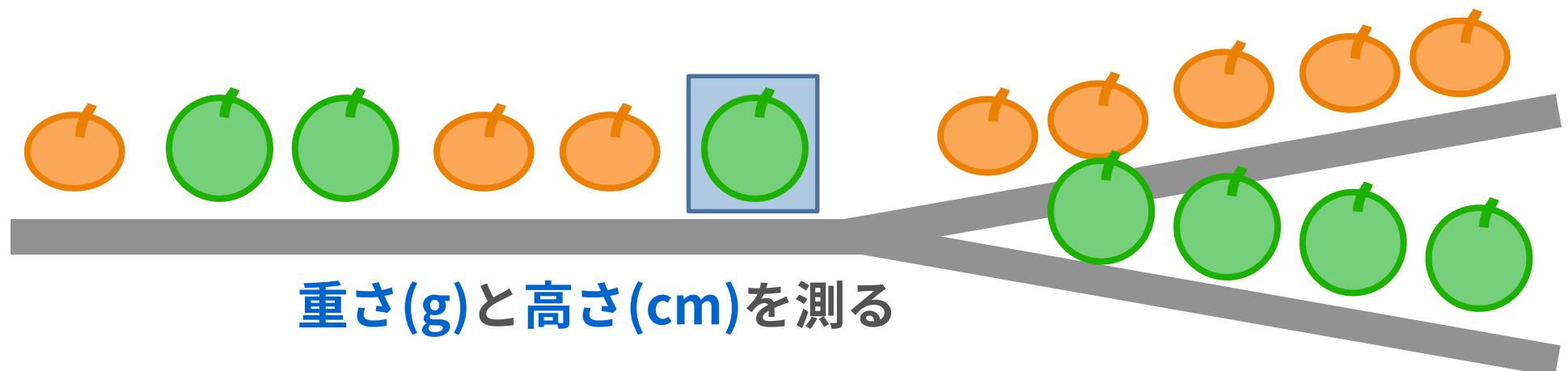
機械学習は「データを予測に変える」技術



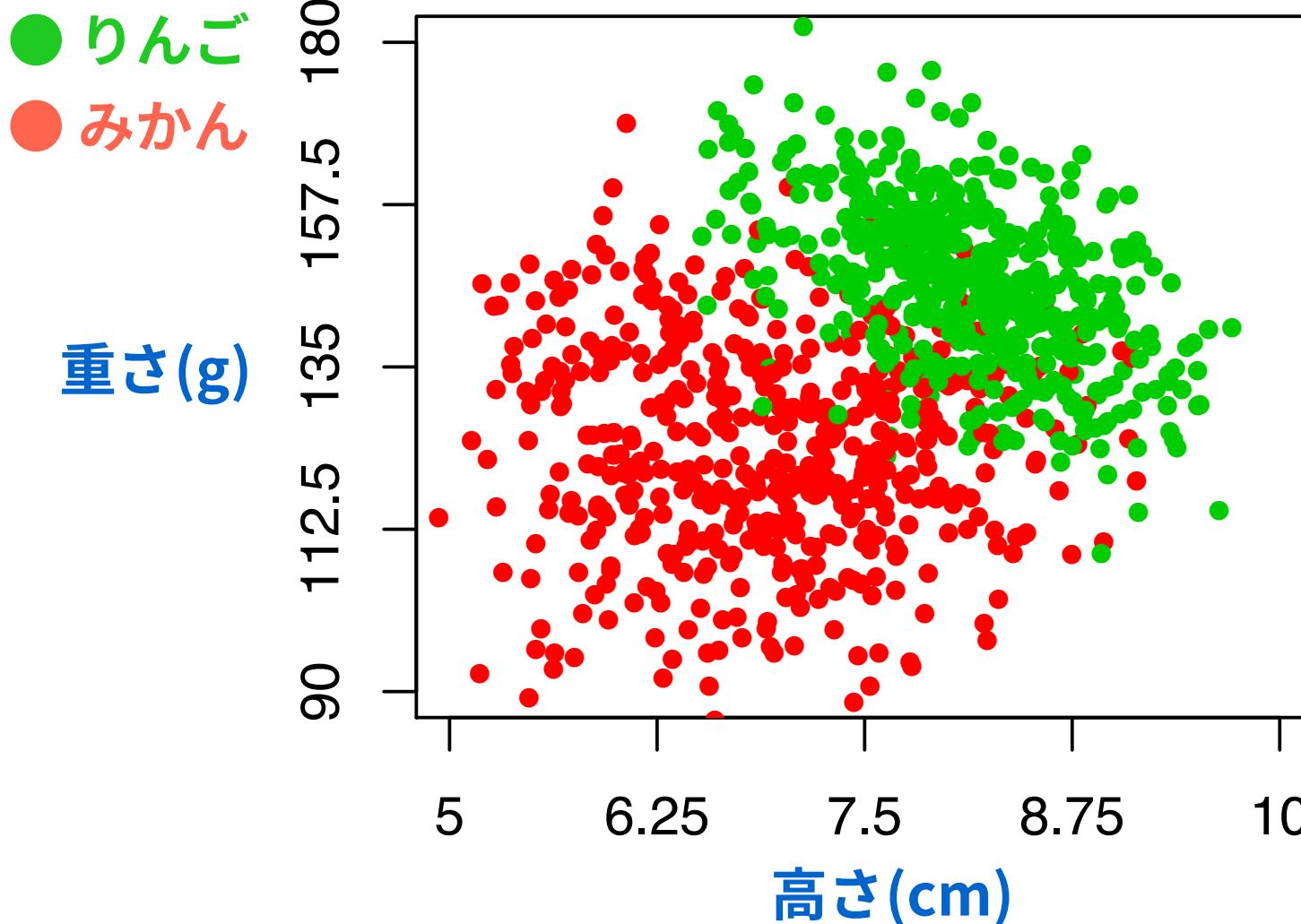
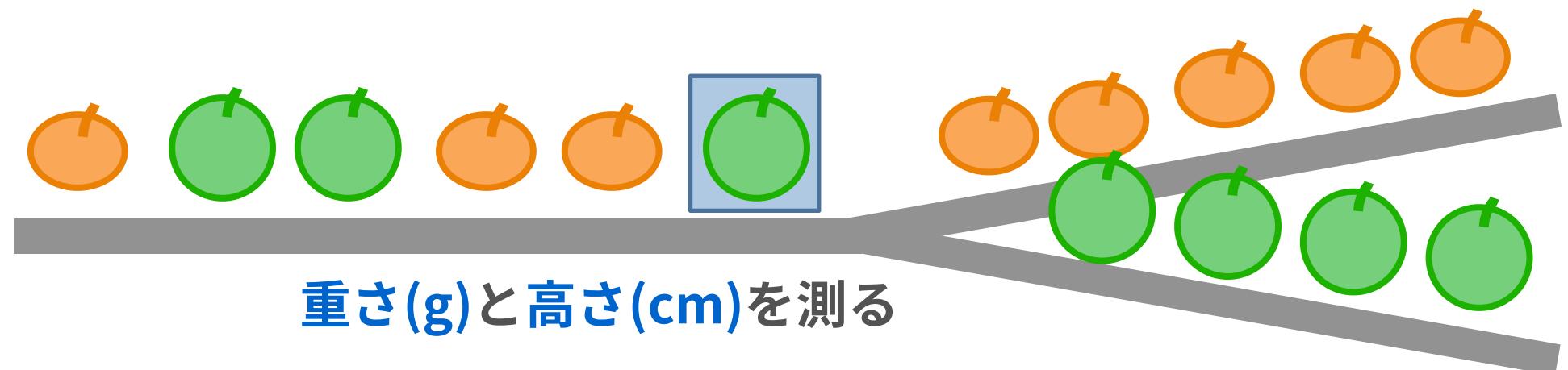
機械学習は「データを予測に変える」技術



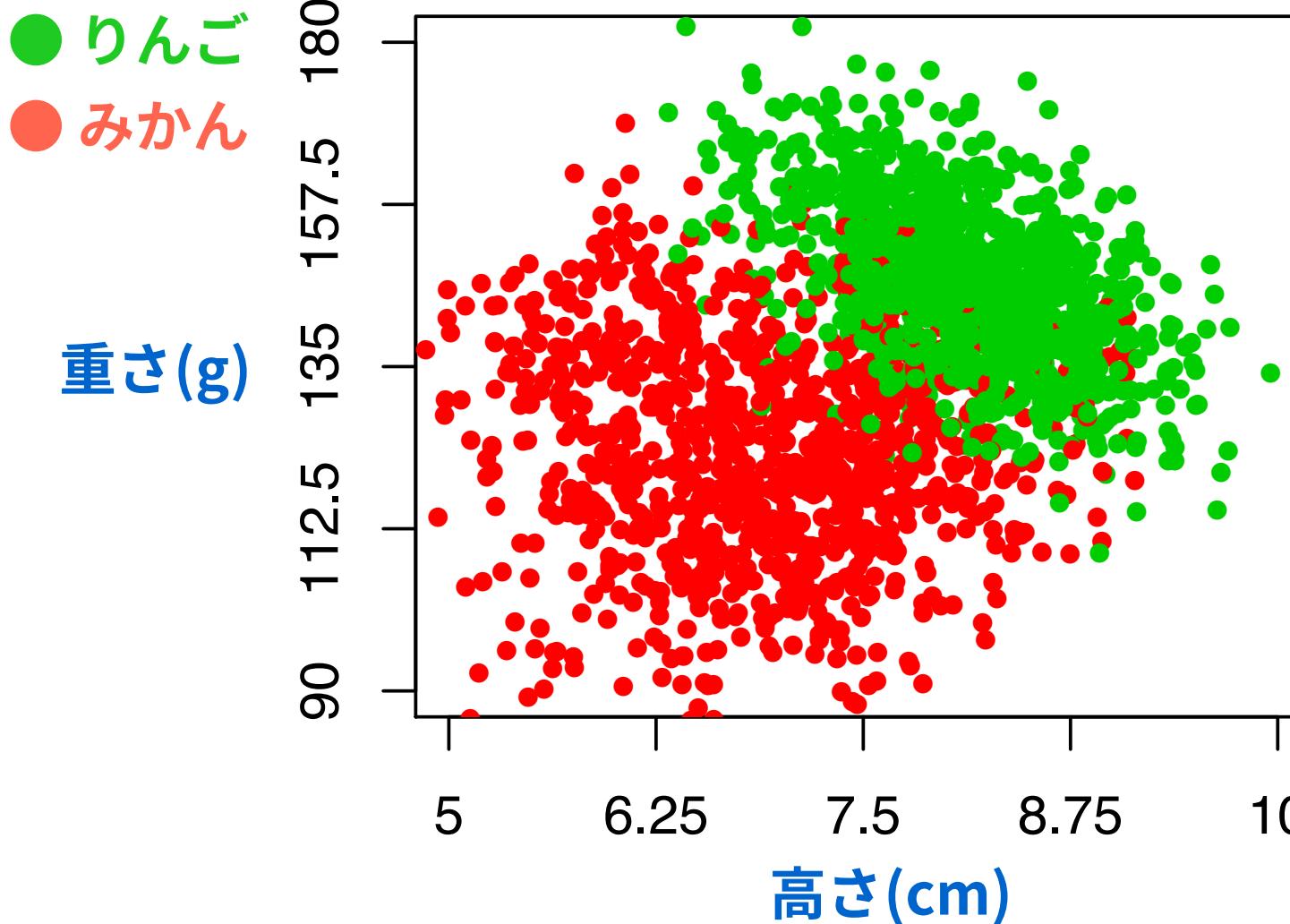
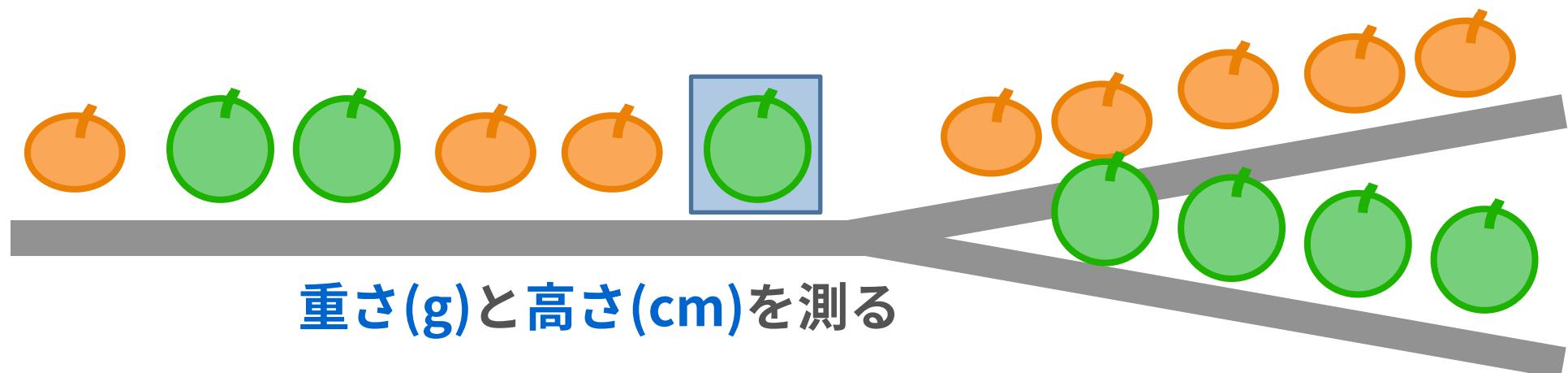
機械学習は「データを予測に変える」技術



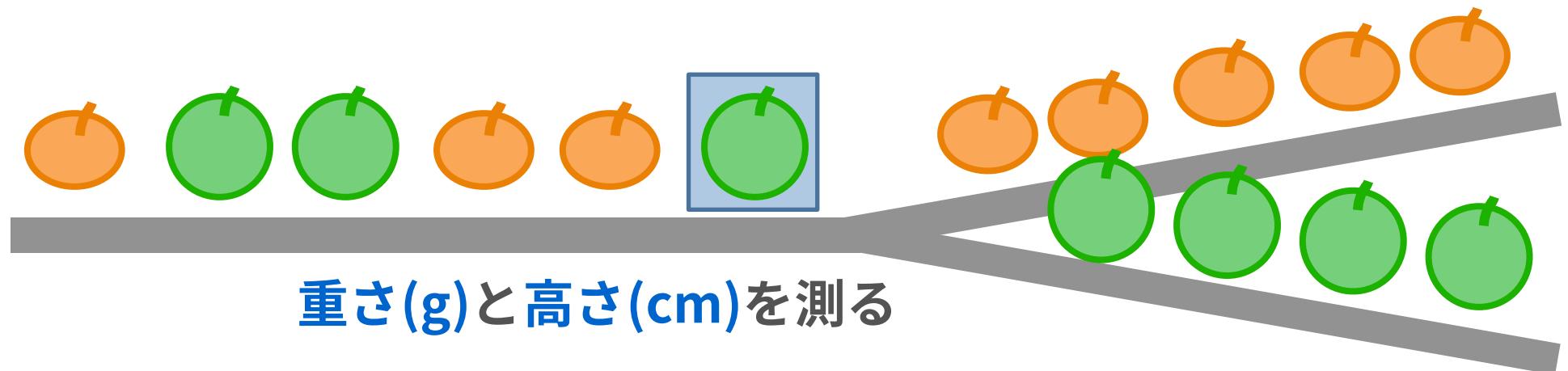
機械学習は「データを予測に変える」技術



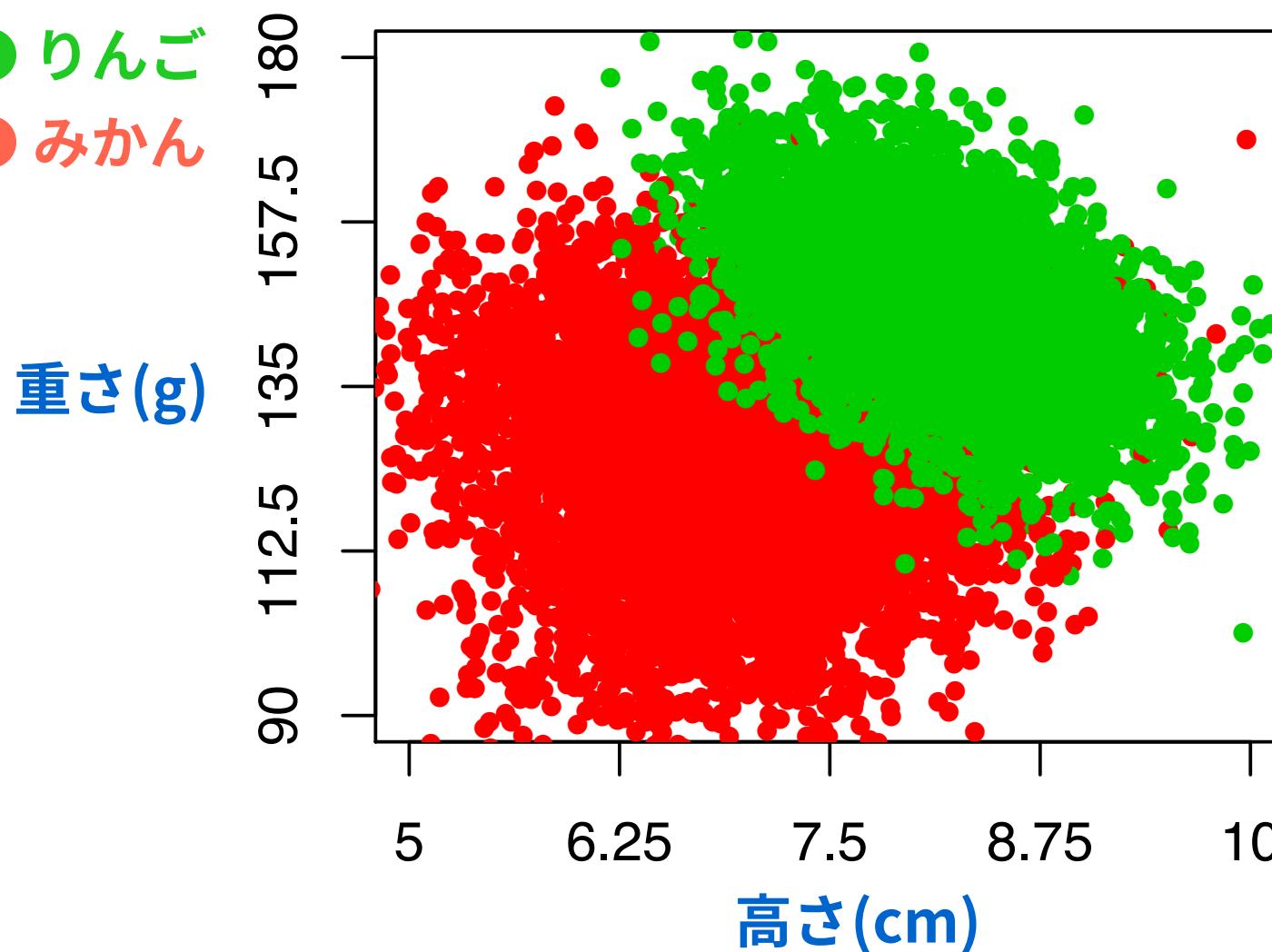
機械学習は「データを予測に変える」技術



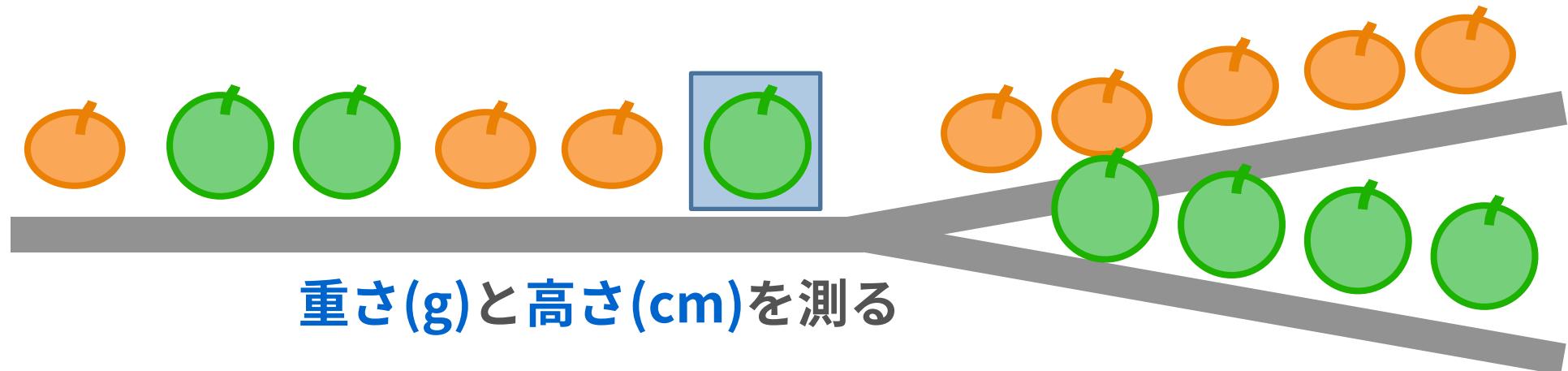
機械学習は「データを予測に変える」技術



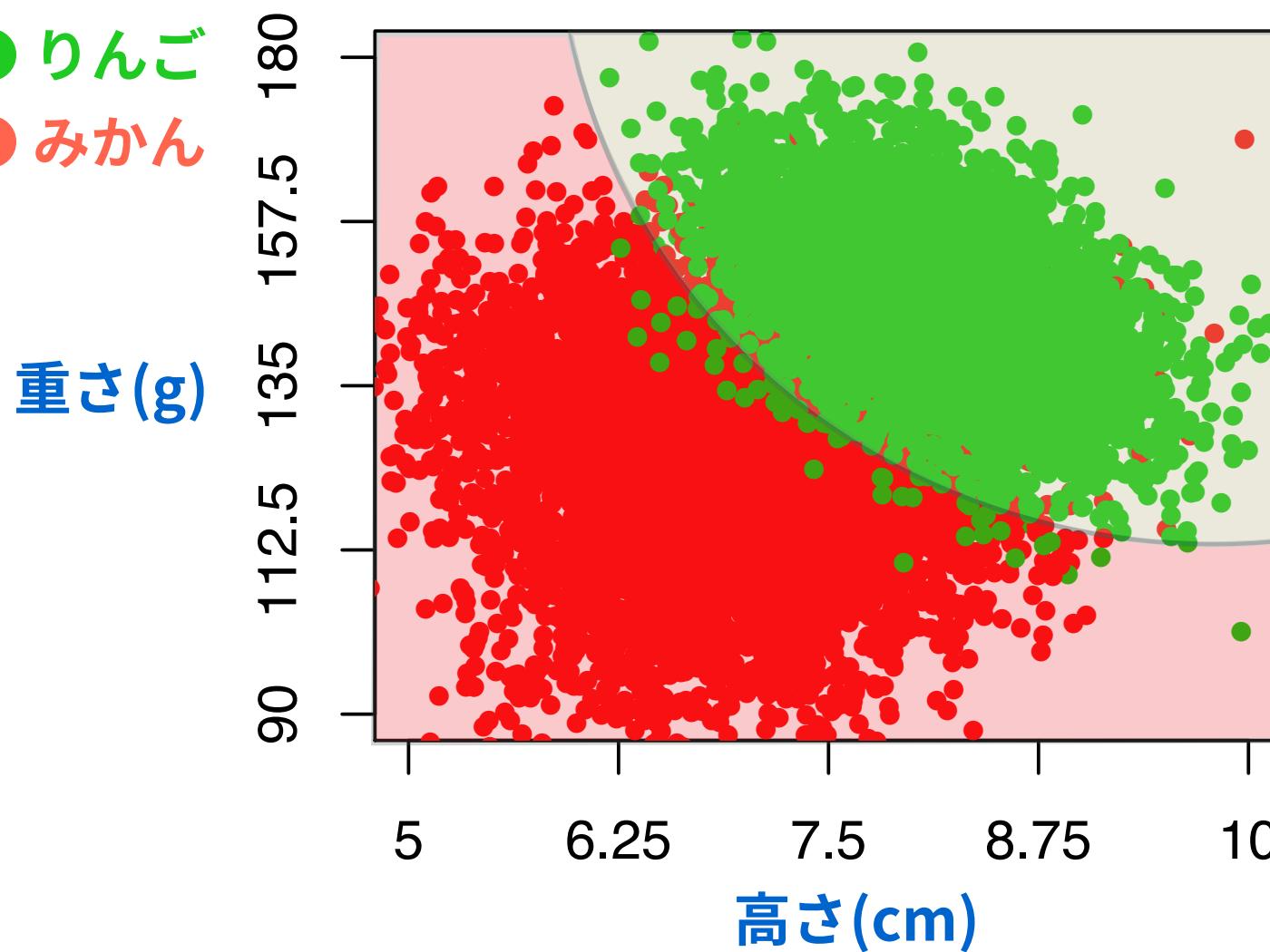
りんご
みかん



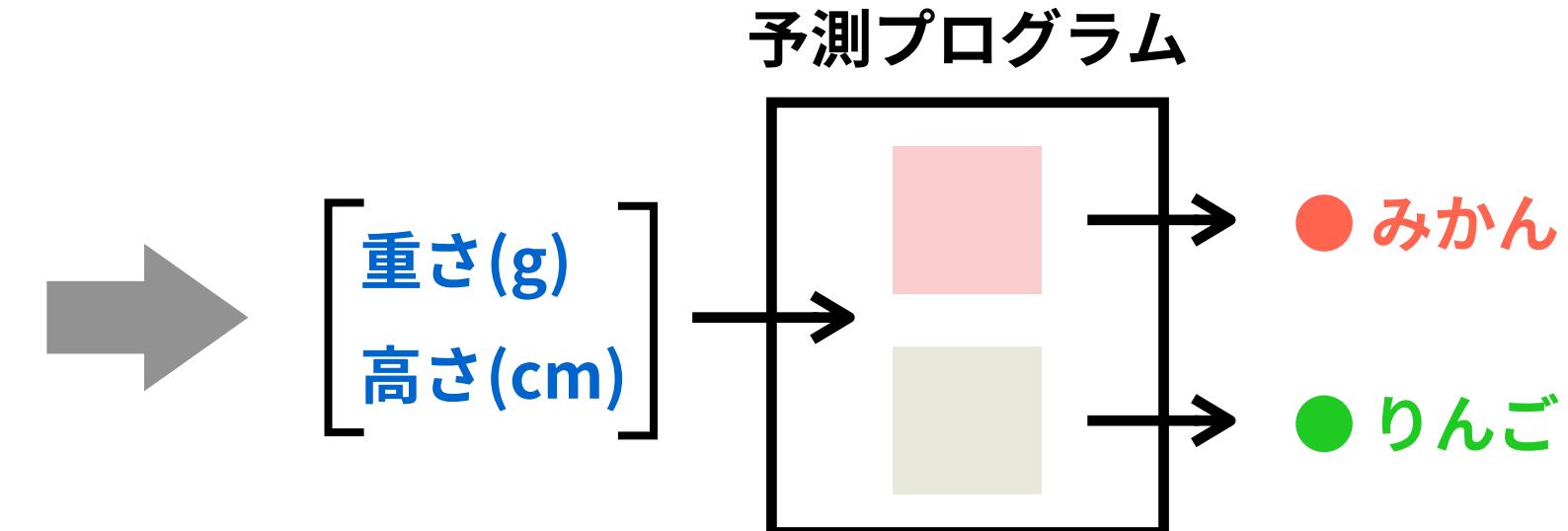
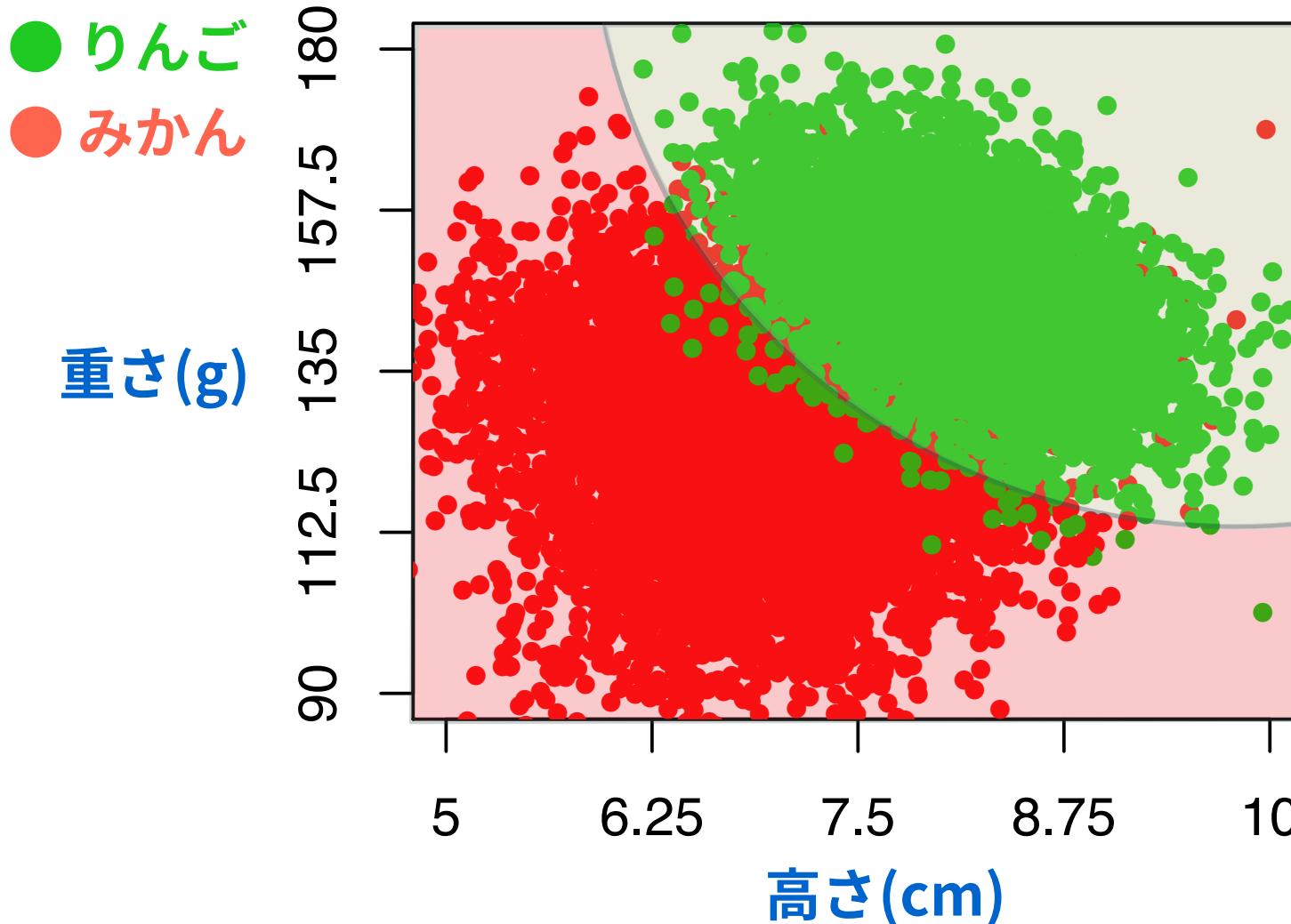
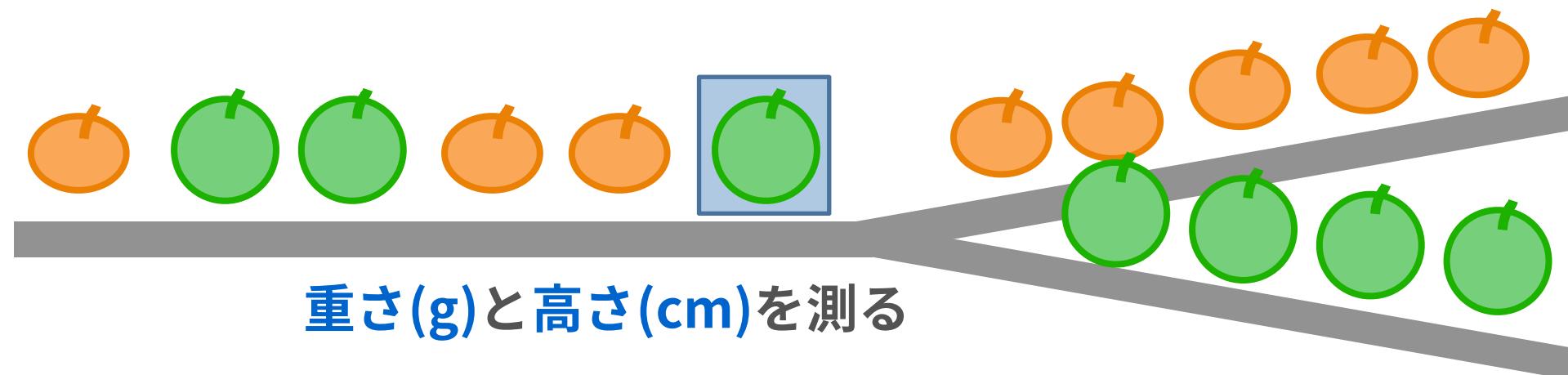
機械学習は「データを予測に変える」技術



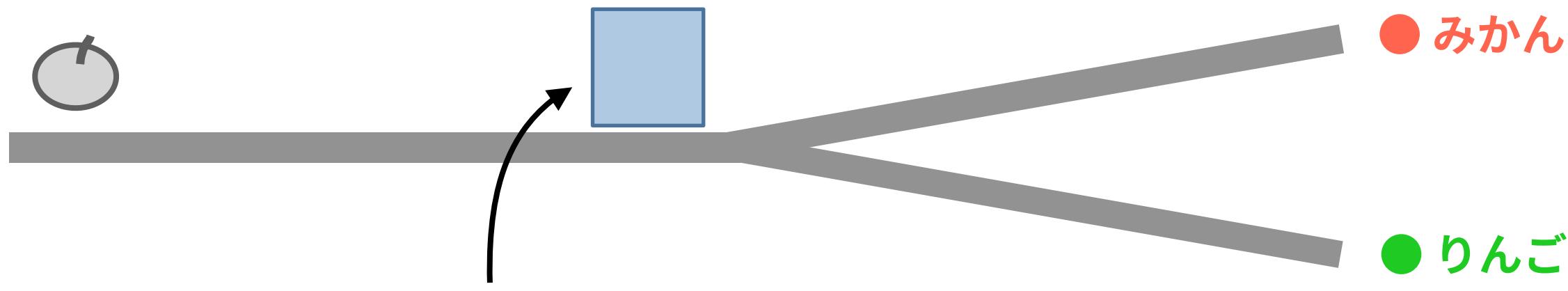
りんご
みかん



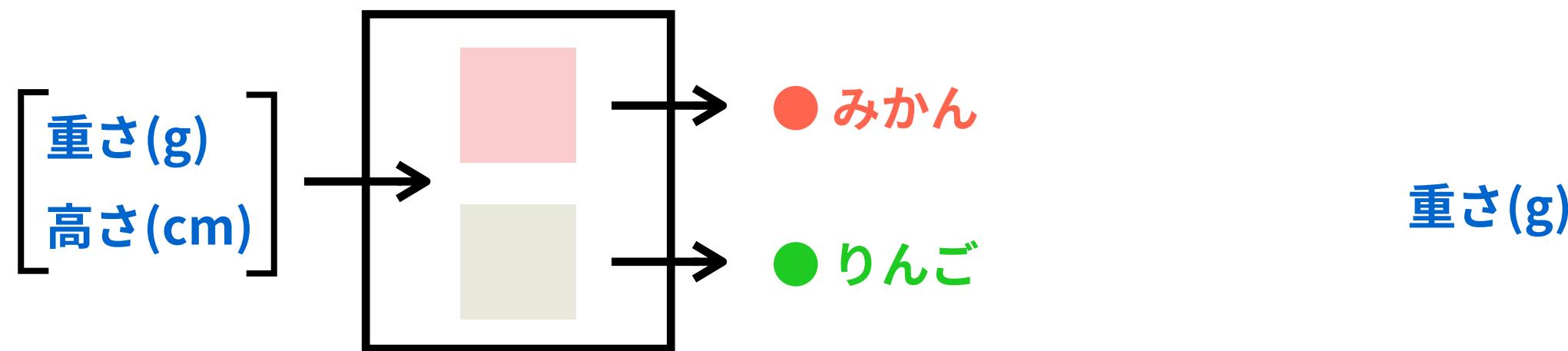
機械学習は「データを予測に変える」技術



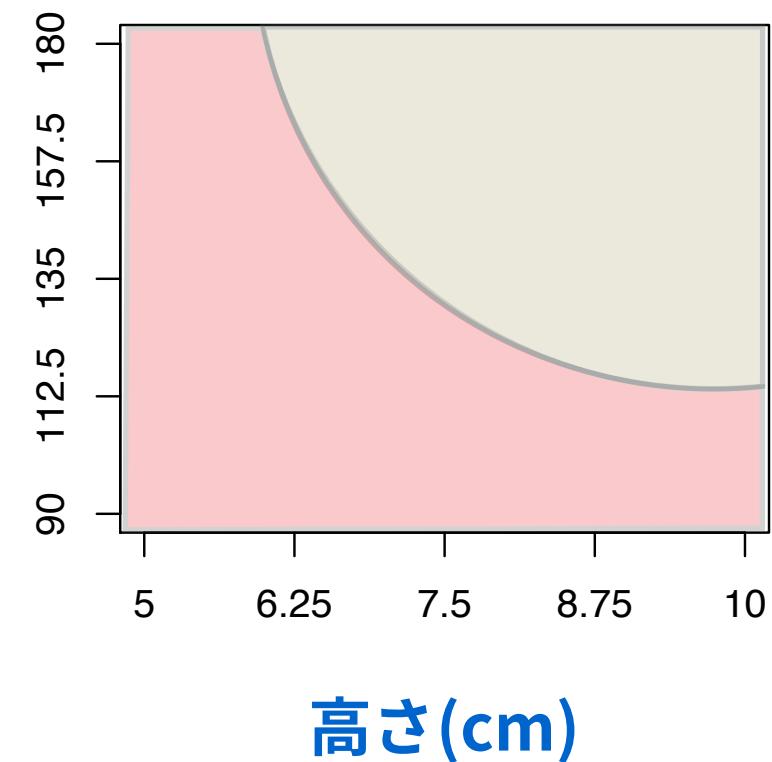
機械学習は「データを予測に変える」技術



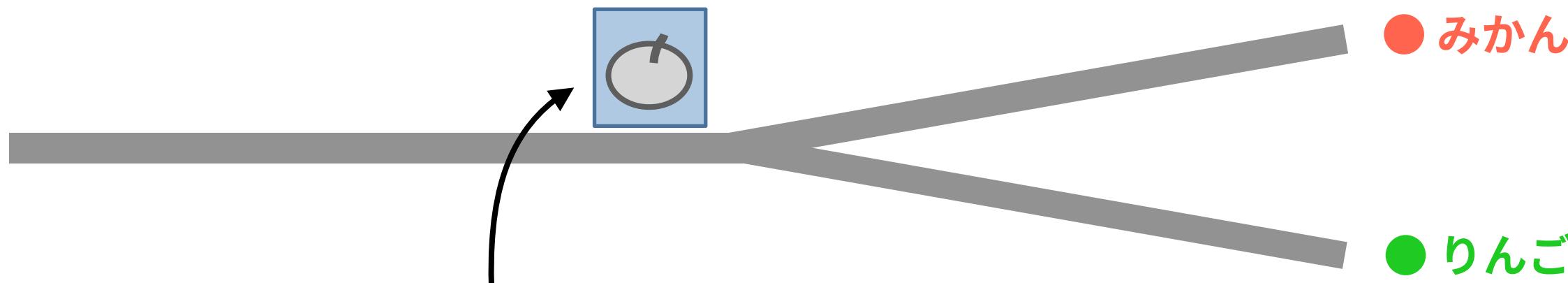
見本データから作っておいた予測プログラム



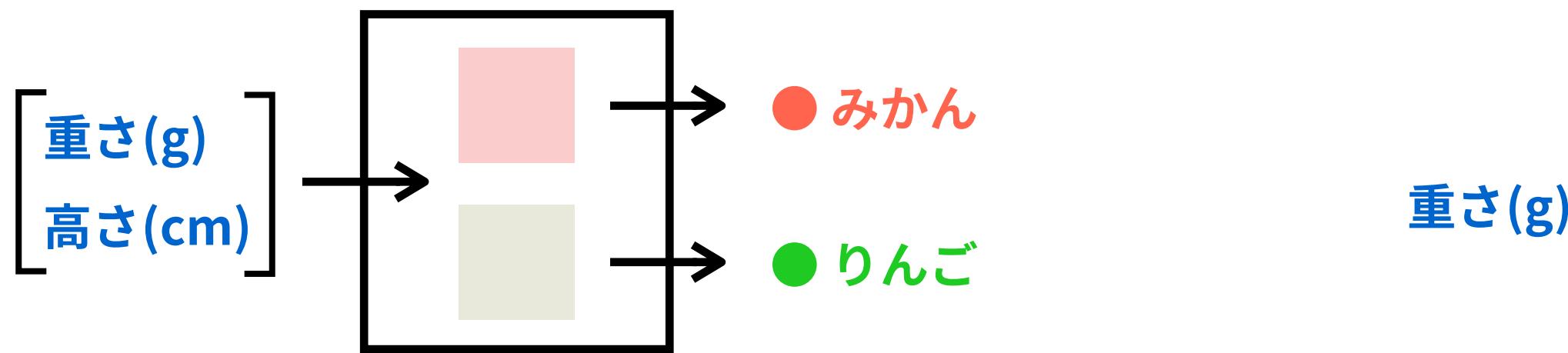
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



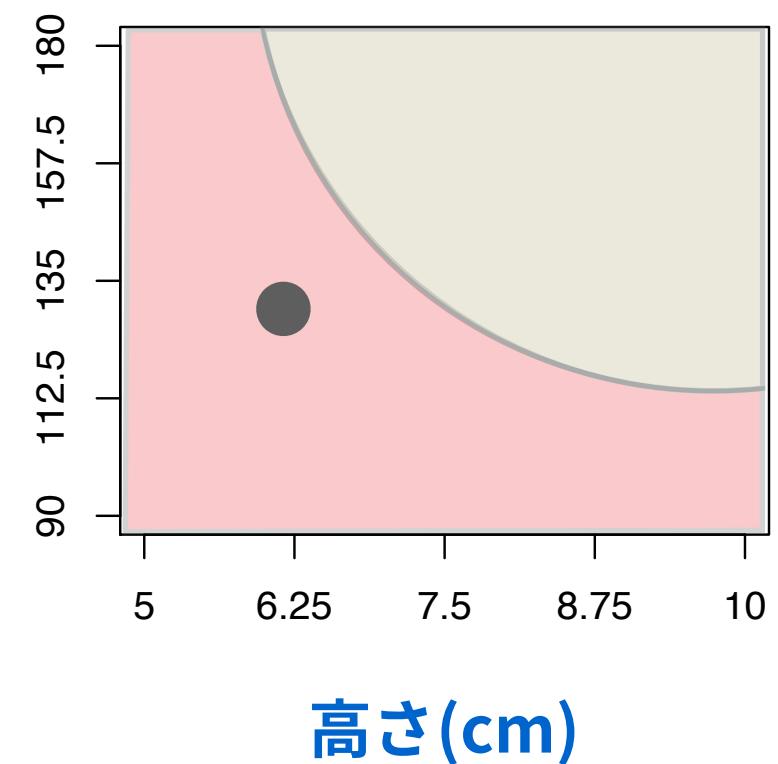
機械学習は「データを予測に変える」技術



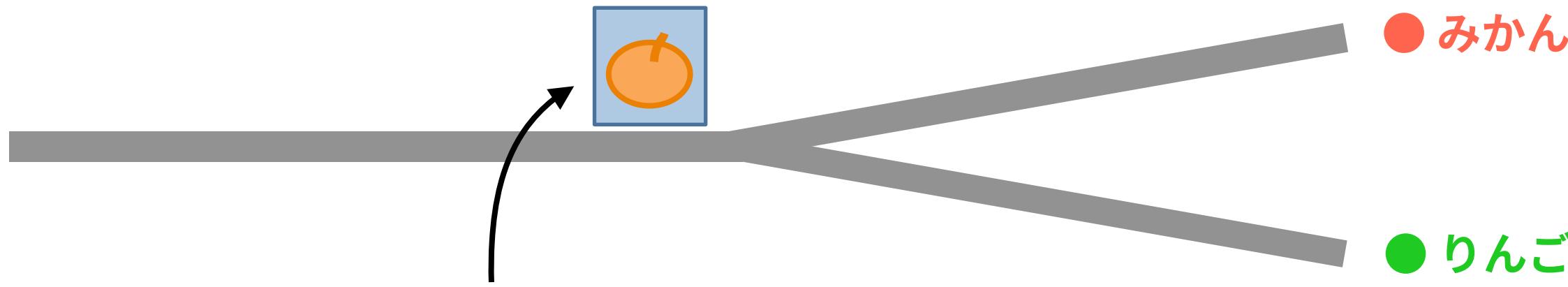
見本データから作っておいた予測プログラム



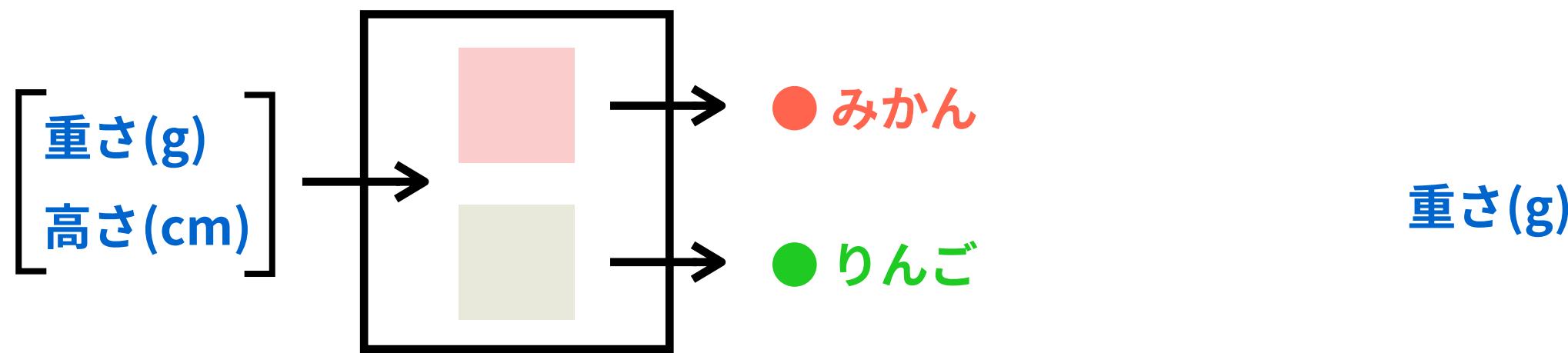
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



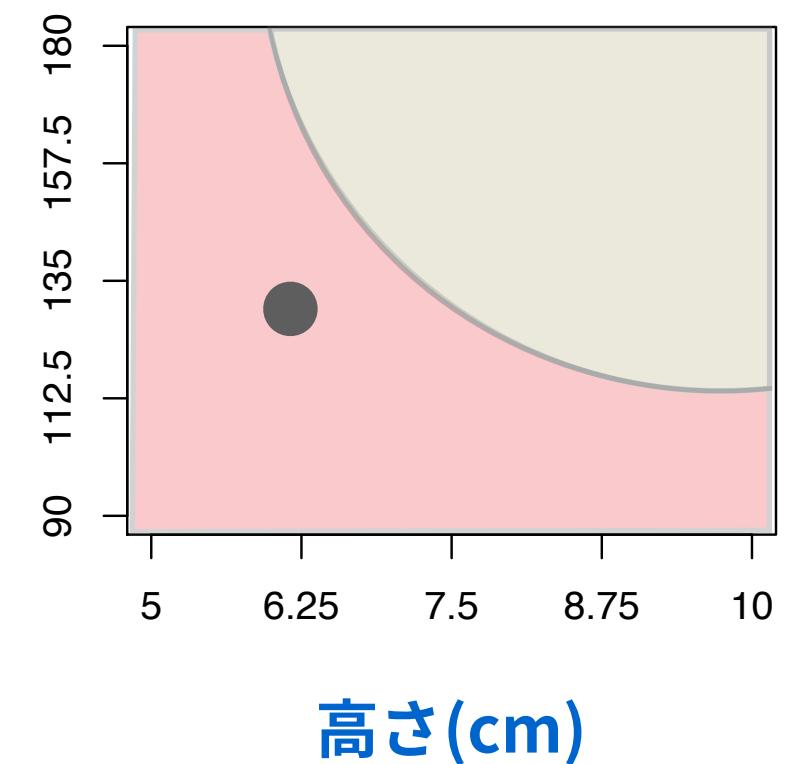
機械学習は「データを予測に変える」技術



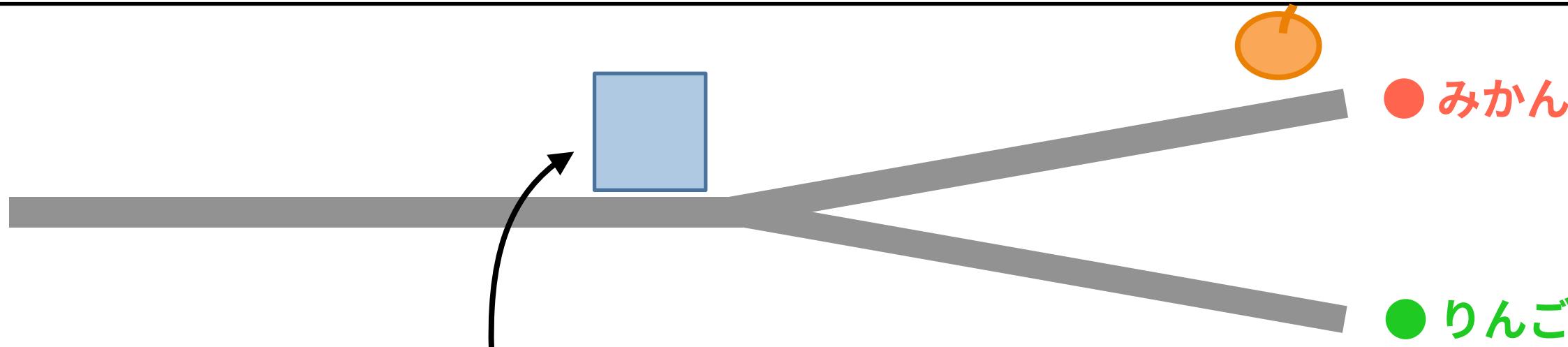
見本データから作っておいた予測プログラム



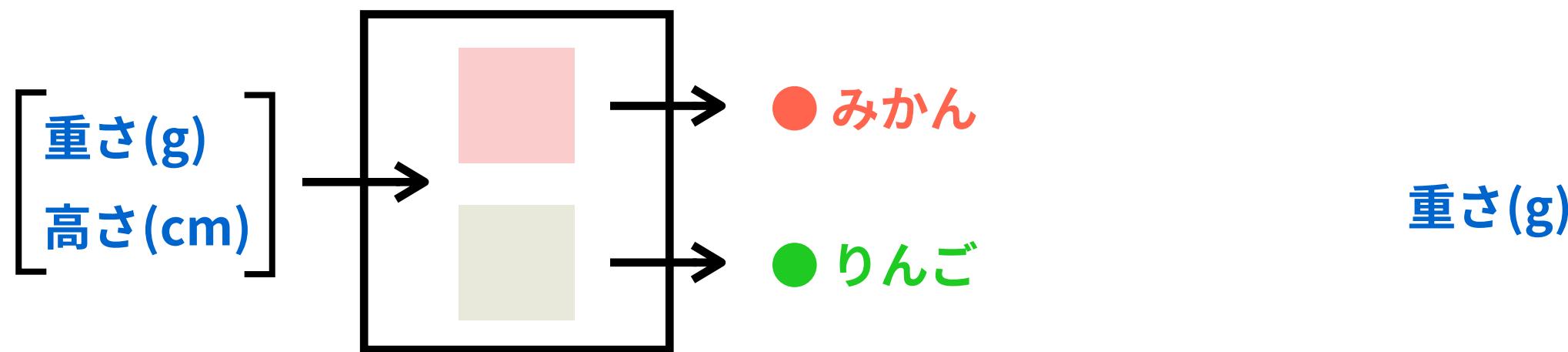
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



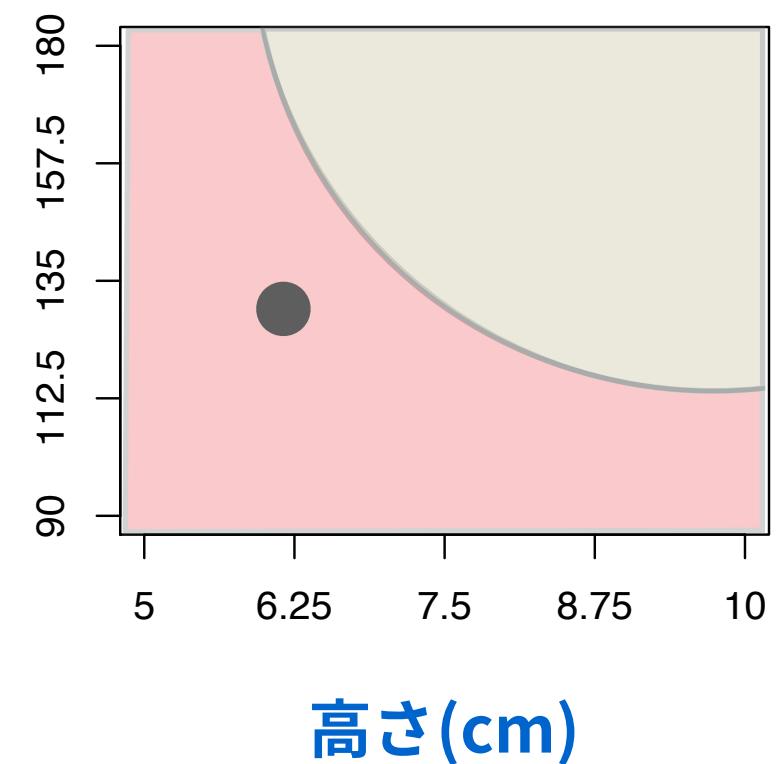
機械学習は「データを予測に変える」技術



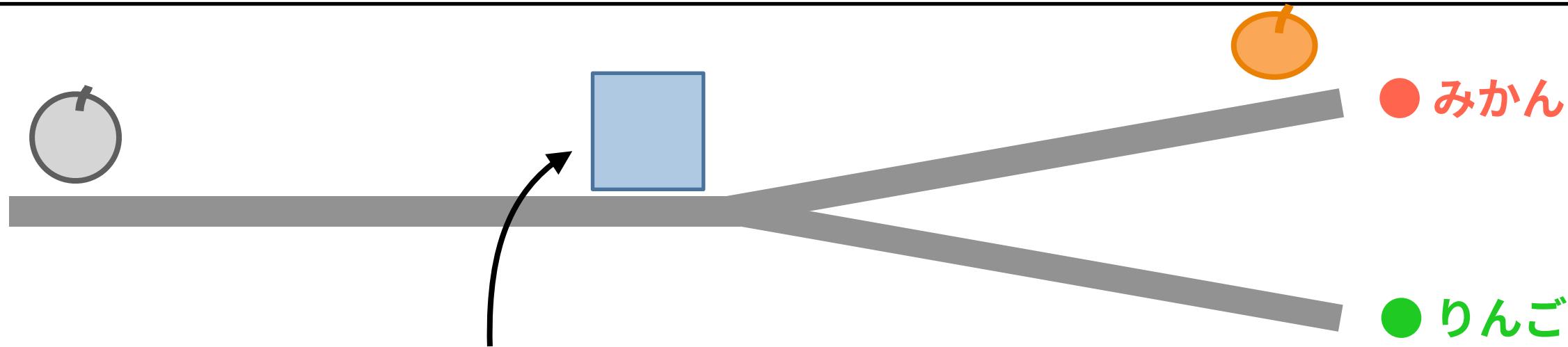
見本データから作っておいた予測プログラム



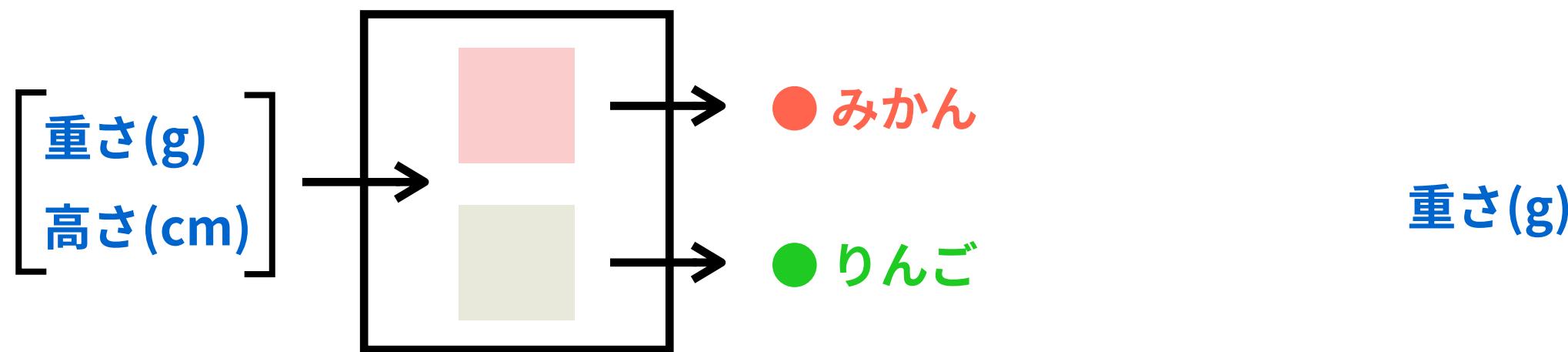
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



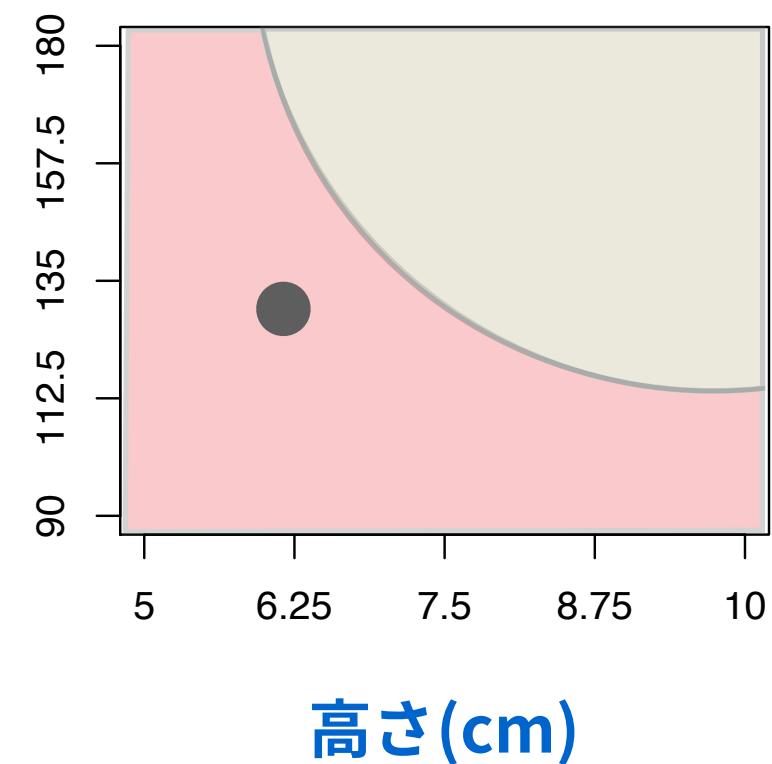
機械学習は「データを予測に変える」技術



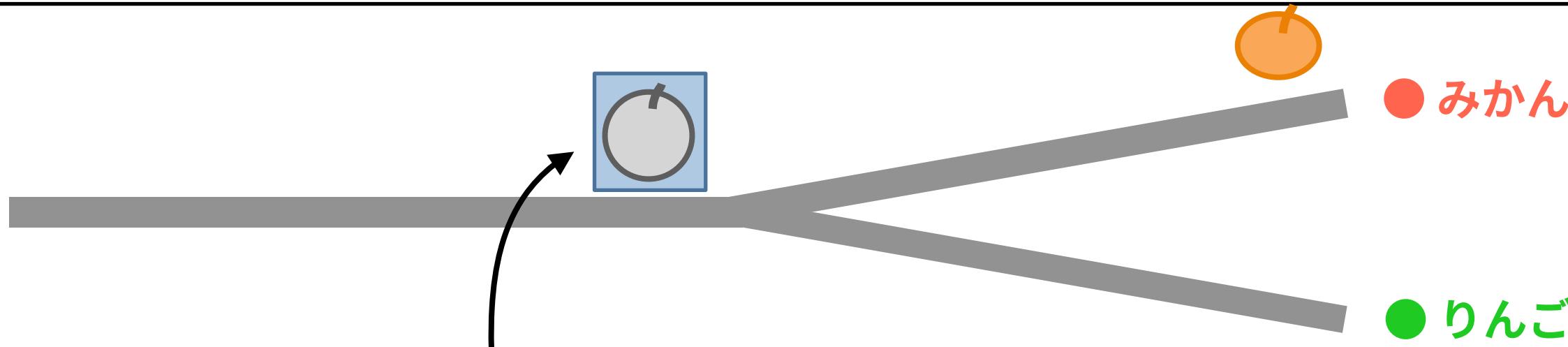
見本データから作っておいた予測プログラム



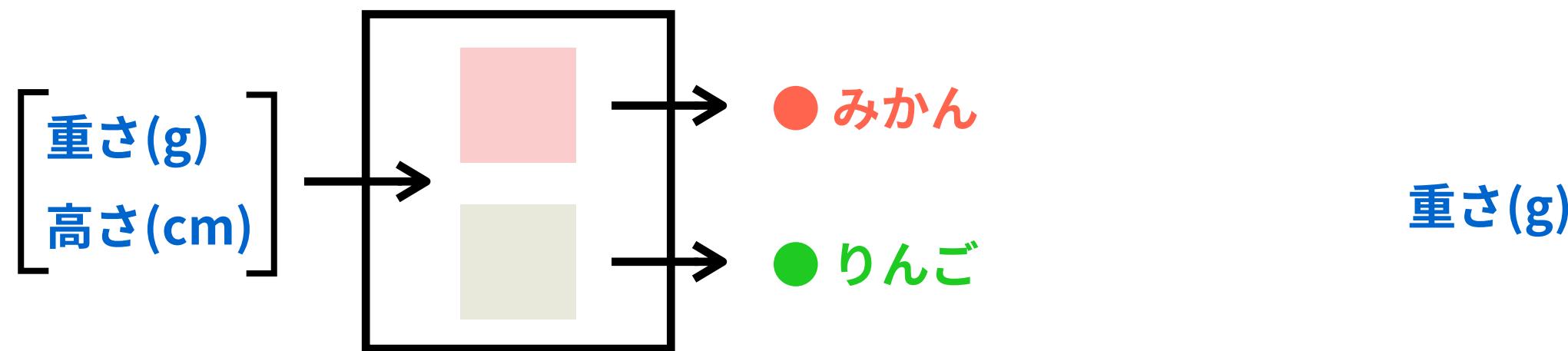
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



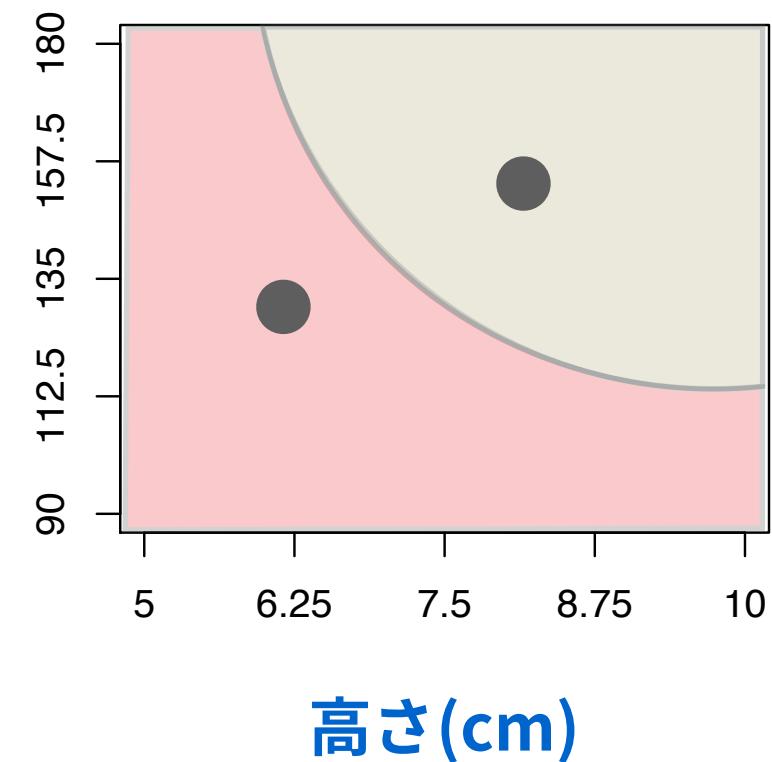
機械学習は「データを予測に変える」技術



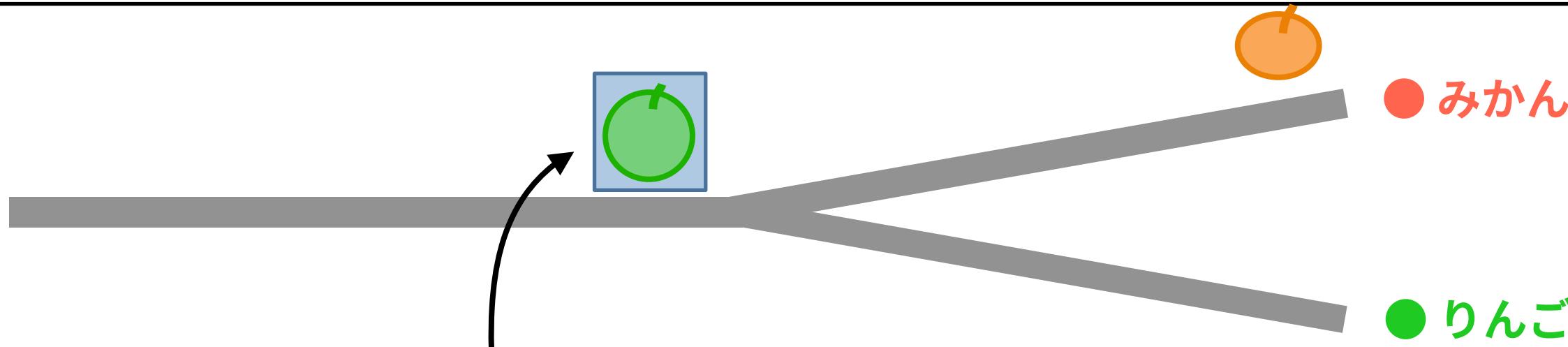
見本データから作っておいた予測プログラム



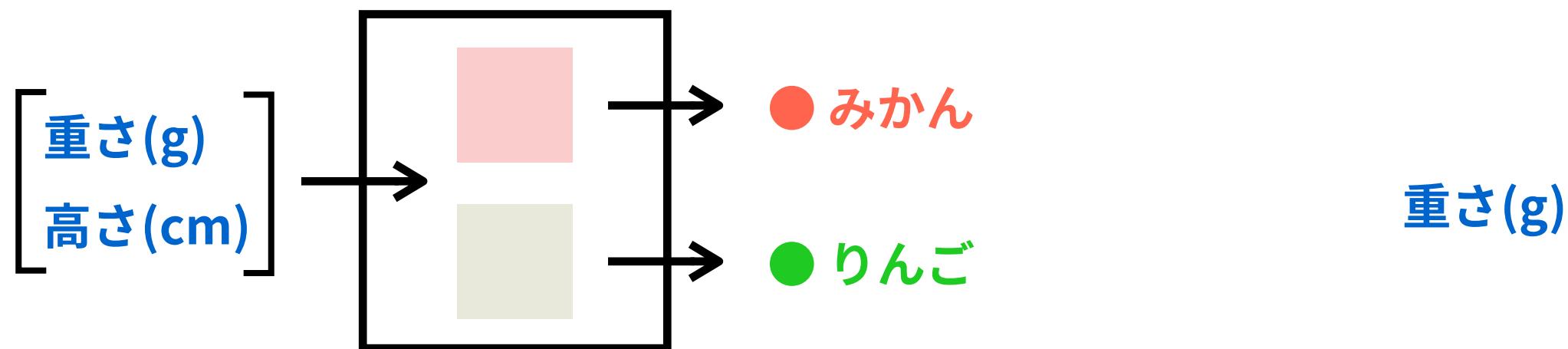
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！



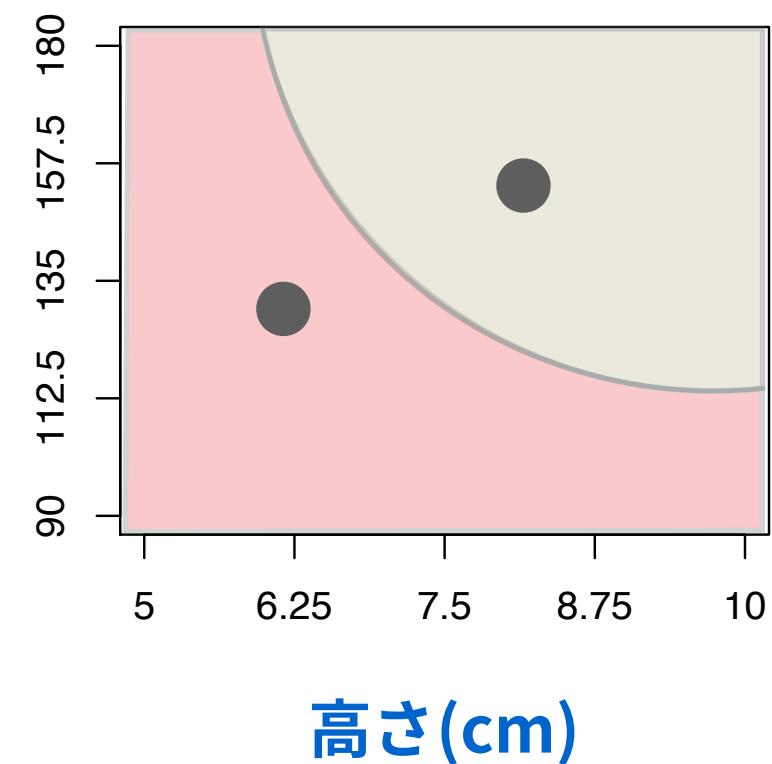
機械学習は「データを予測に変える」技術



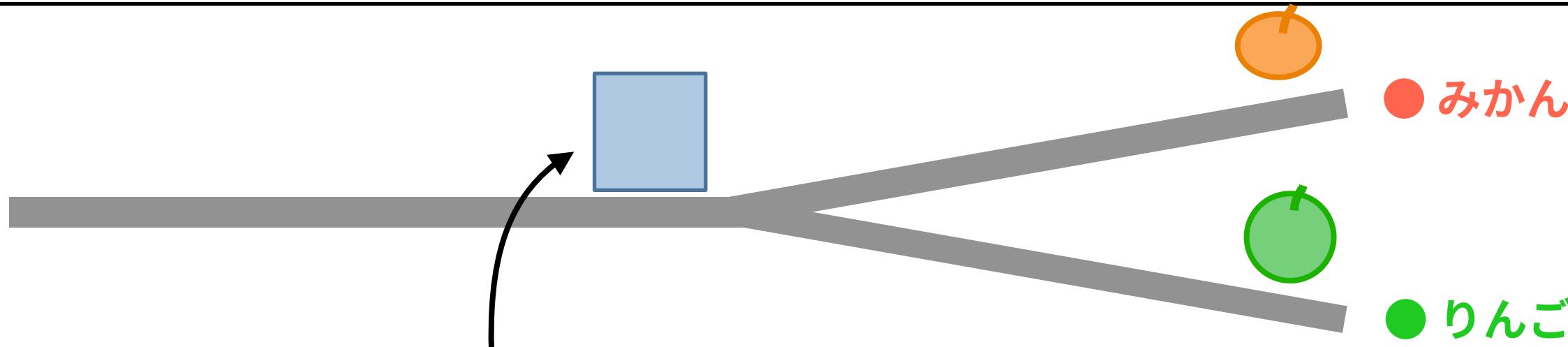
見本データから作っておいた予測プログラム



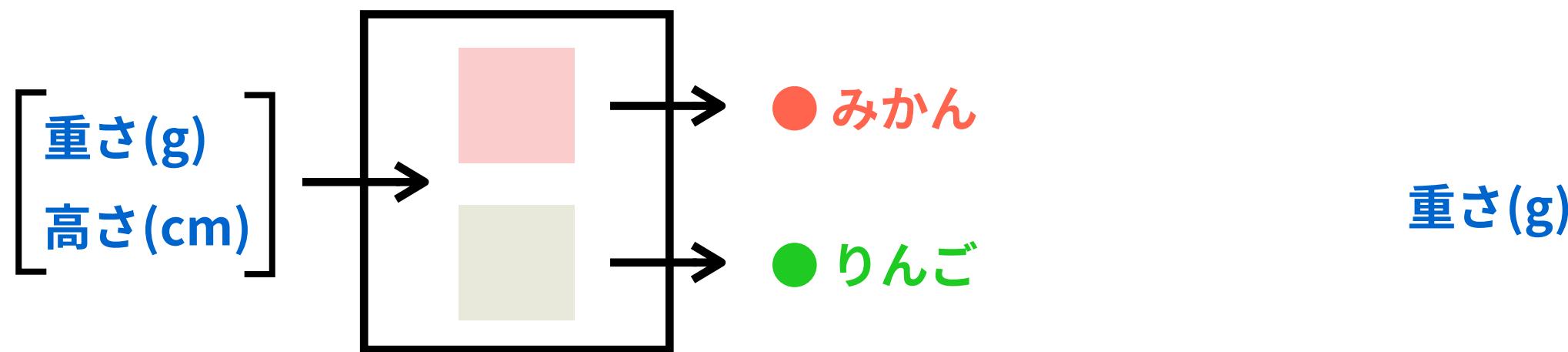
予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」 を予測することができる！



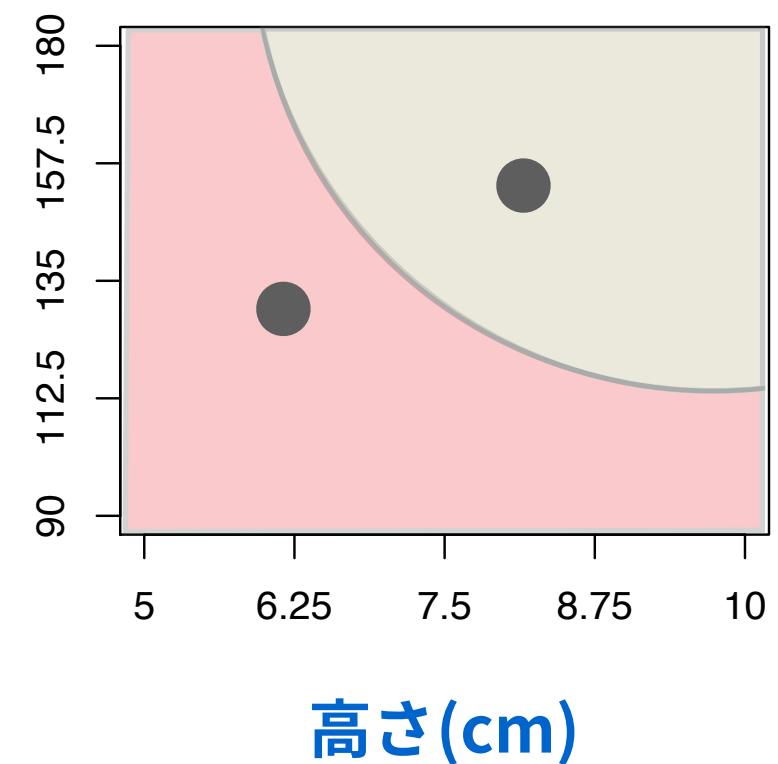
機械学習は「データを予測に変える」技術



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

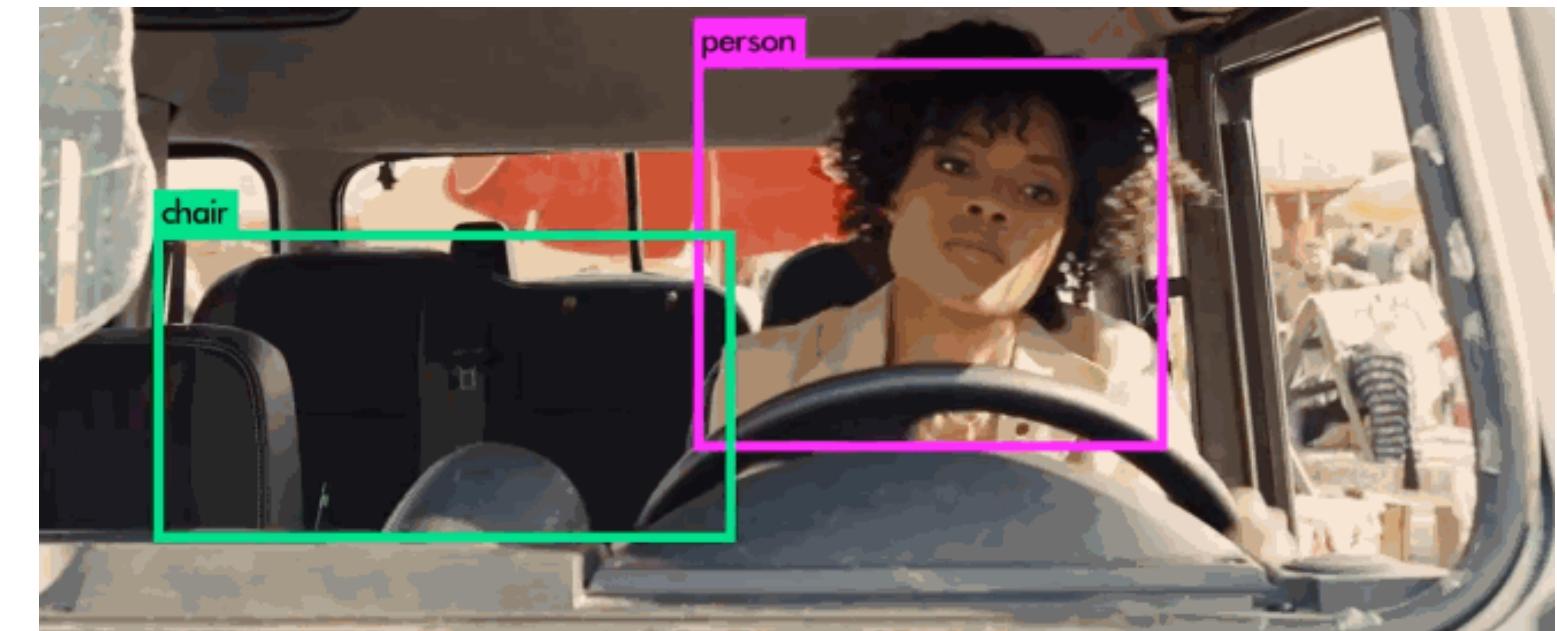
の入力と出力の関係がよく分からない場合でも、

たくさんの入出力の見本データによって間接的に見本を再現するプログラムを作り出す技術



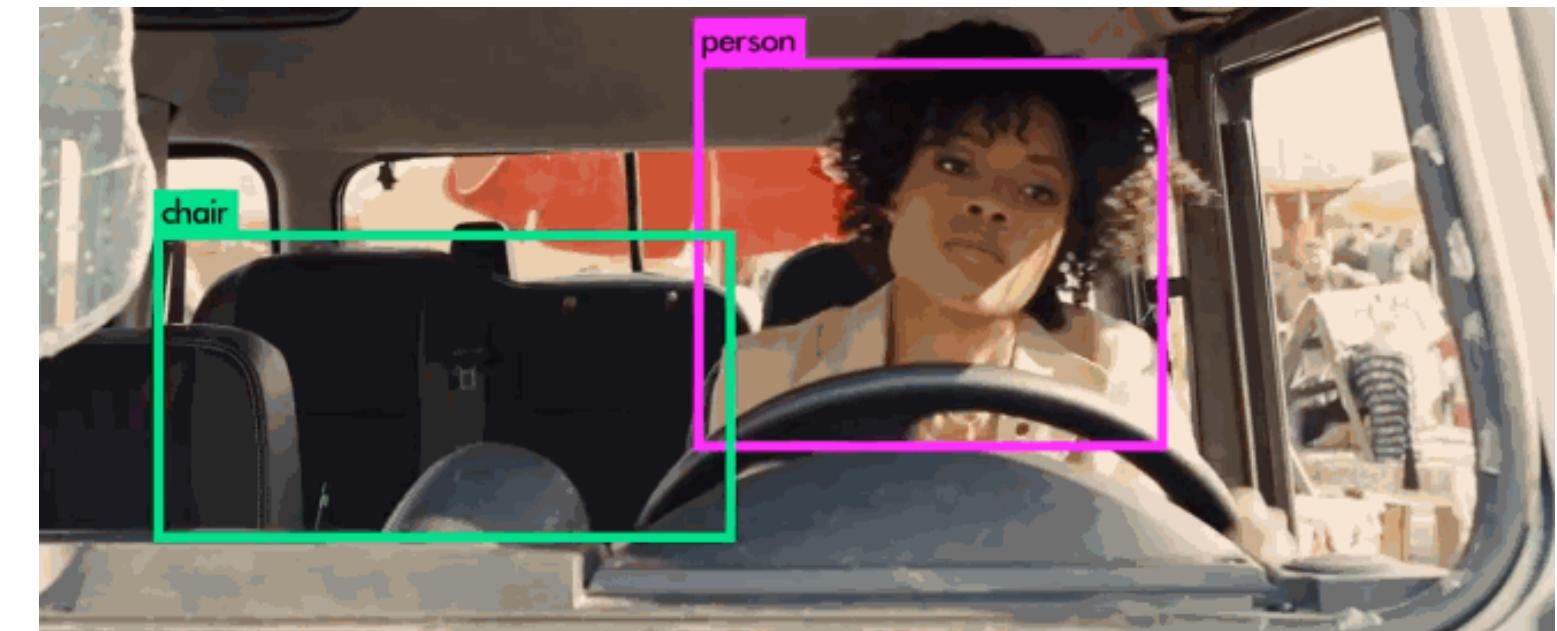
機械学習は「新しい(雑な)コンピュータプログラムの作り方」

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！



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！



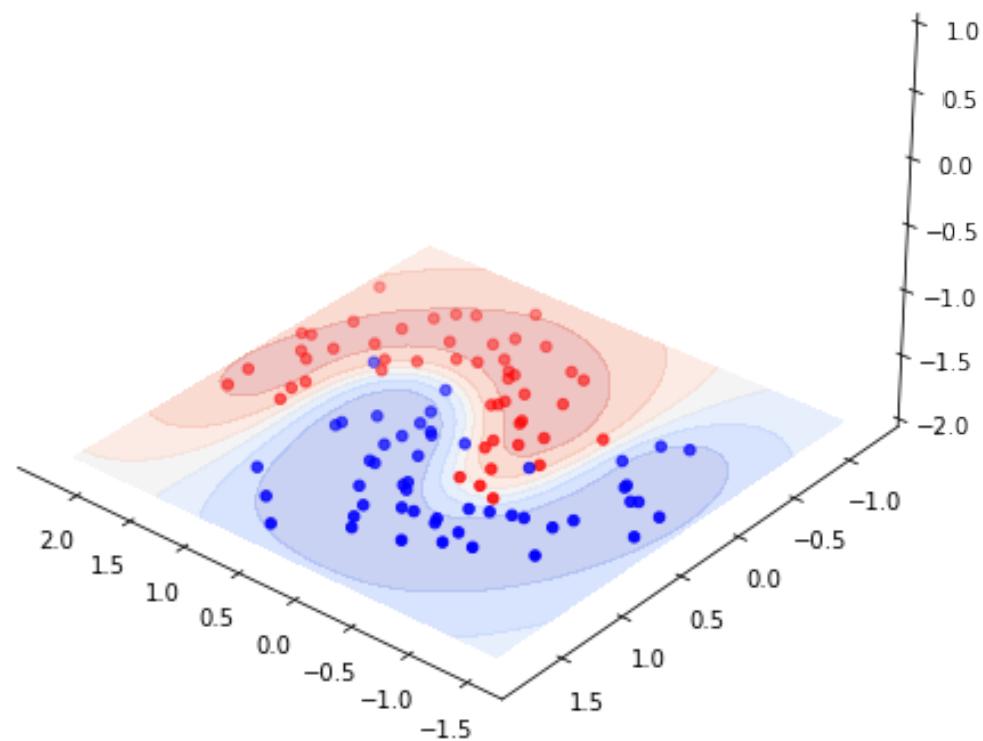
機械学習は「新しい(雑な)コンピュータプログラムの作り方」

機械学習のアルゴリズムはたくさんある。違いは境界線の引き方の方針



仕組みは曲面フィッティングによるデータ内挿

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



仕組みは曲面フィッティングによるデータ内挿

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



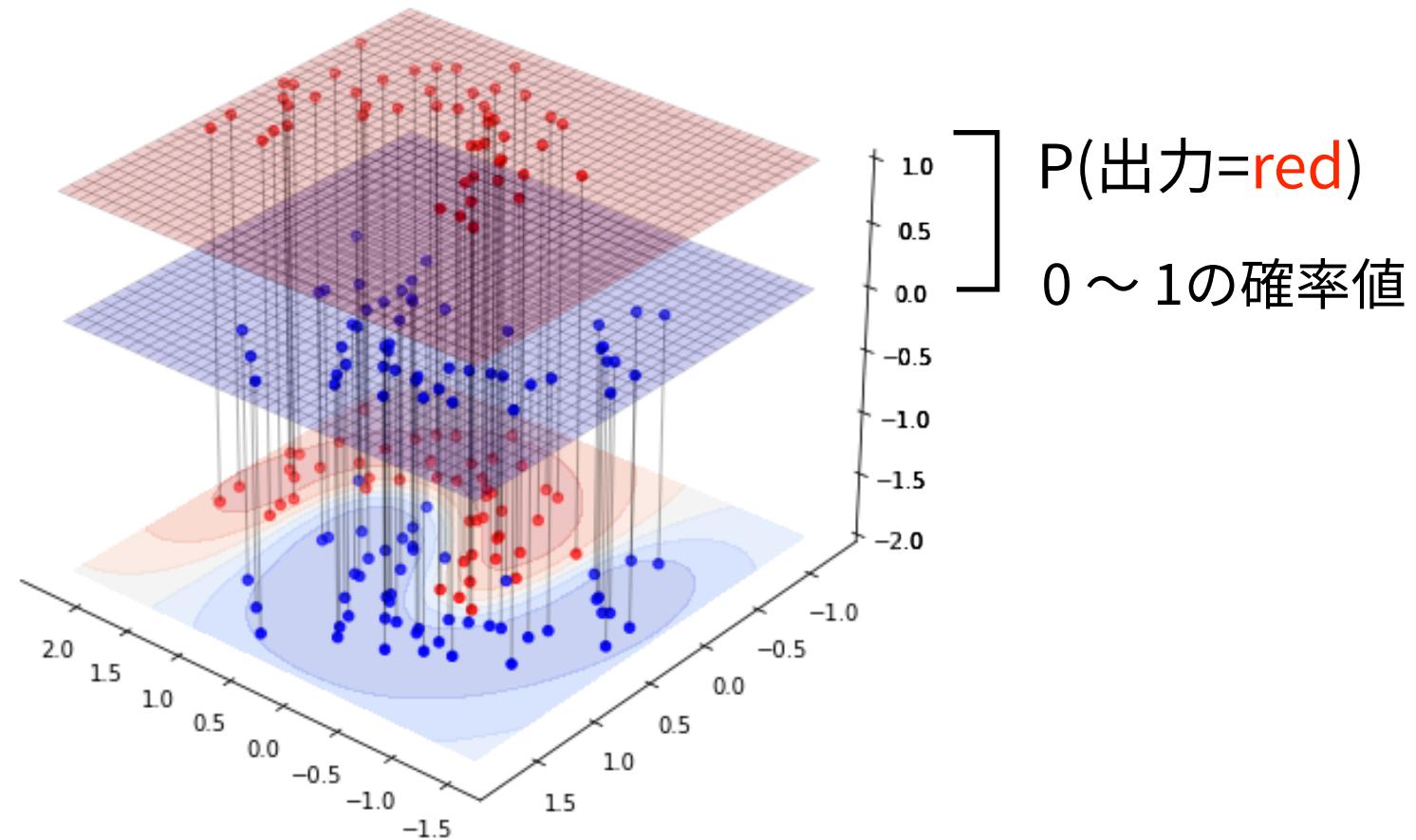
仕組みは曲面フィッティングによるデータ内挿

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



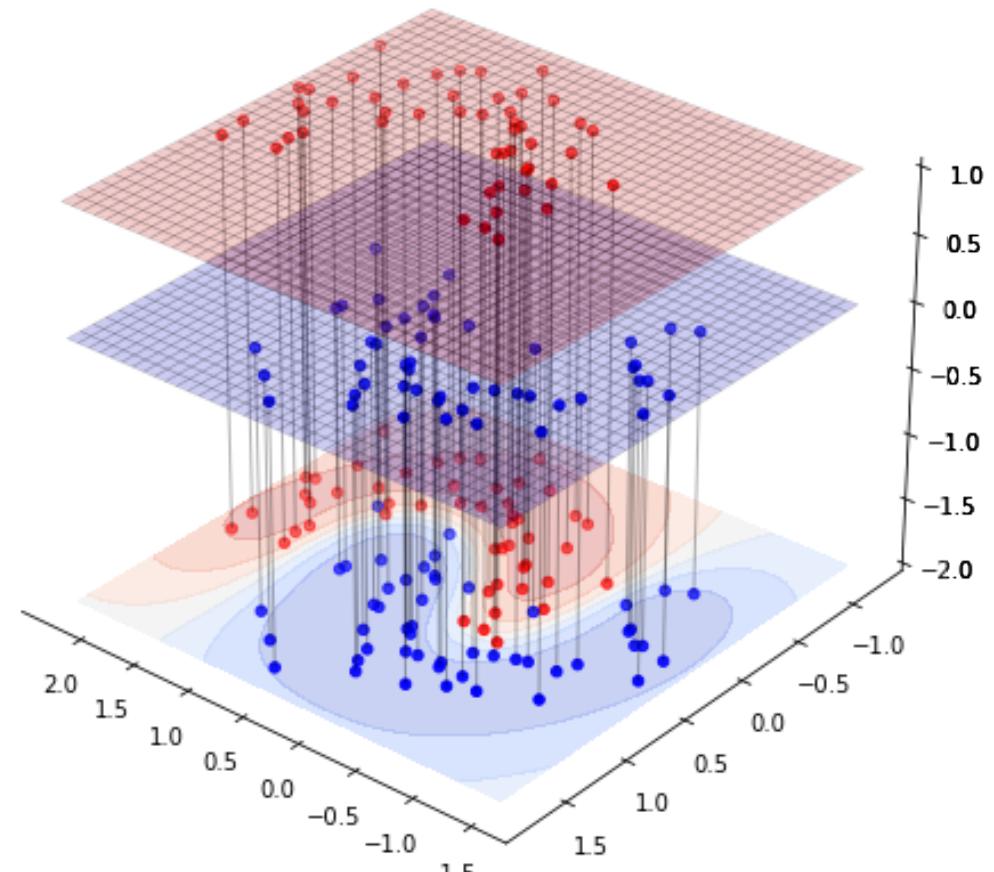
仕組みは曲面フィッティングによるデータ内挿

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

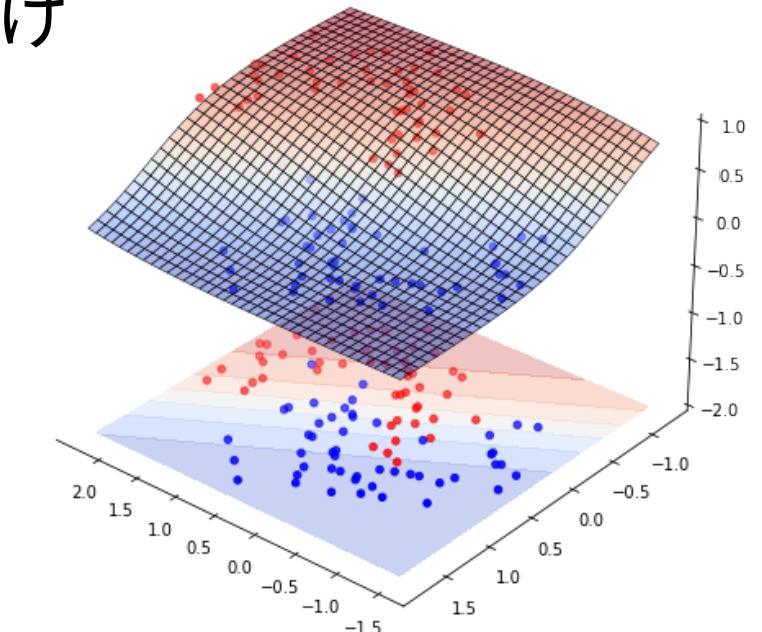


仕組みは曲面フィッティングによるデータ内挿

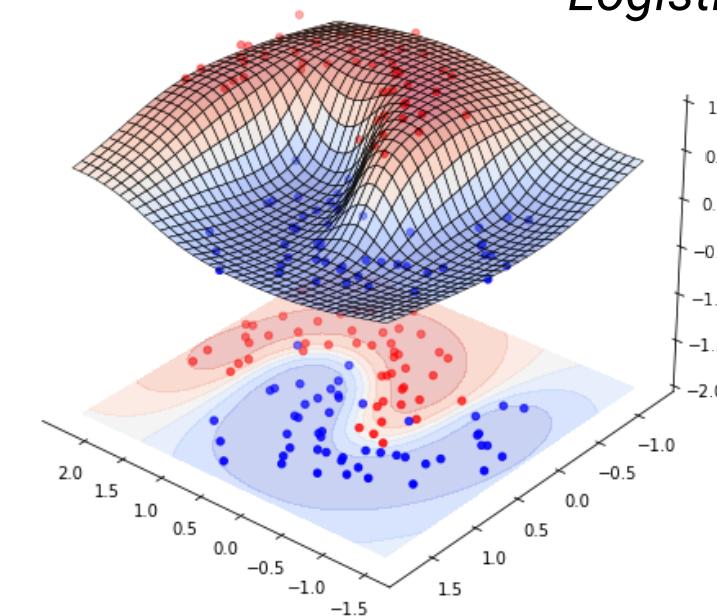
内部原理は「曲面モデル」を点にフィッティングしているだけ



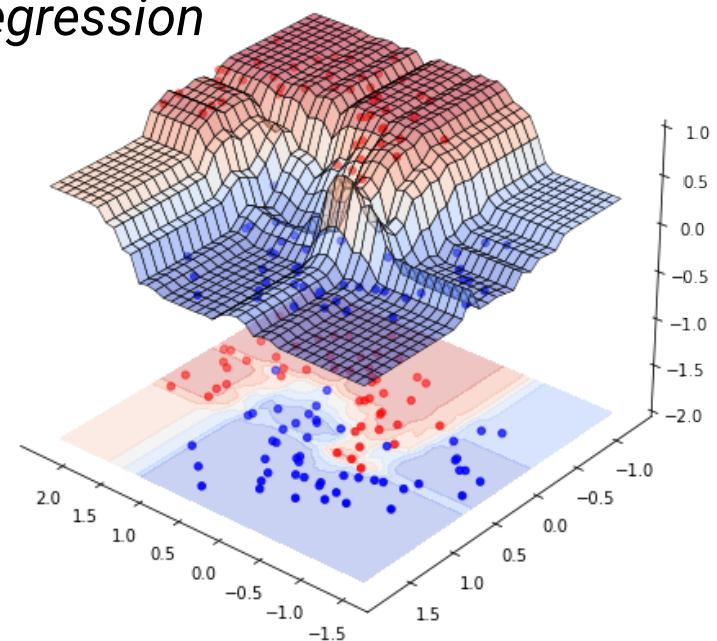
$P(\text{出力}=\text{red})$
0 ~ 1の確率値



Logistic Regression



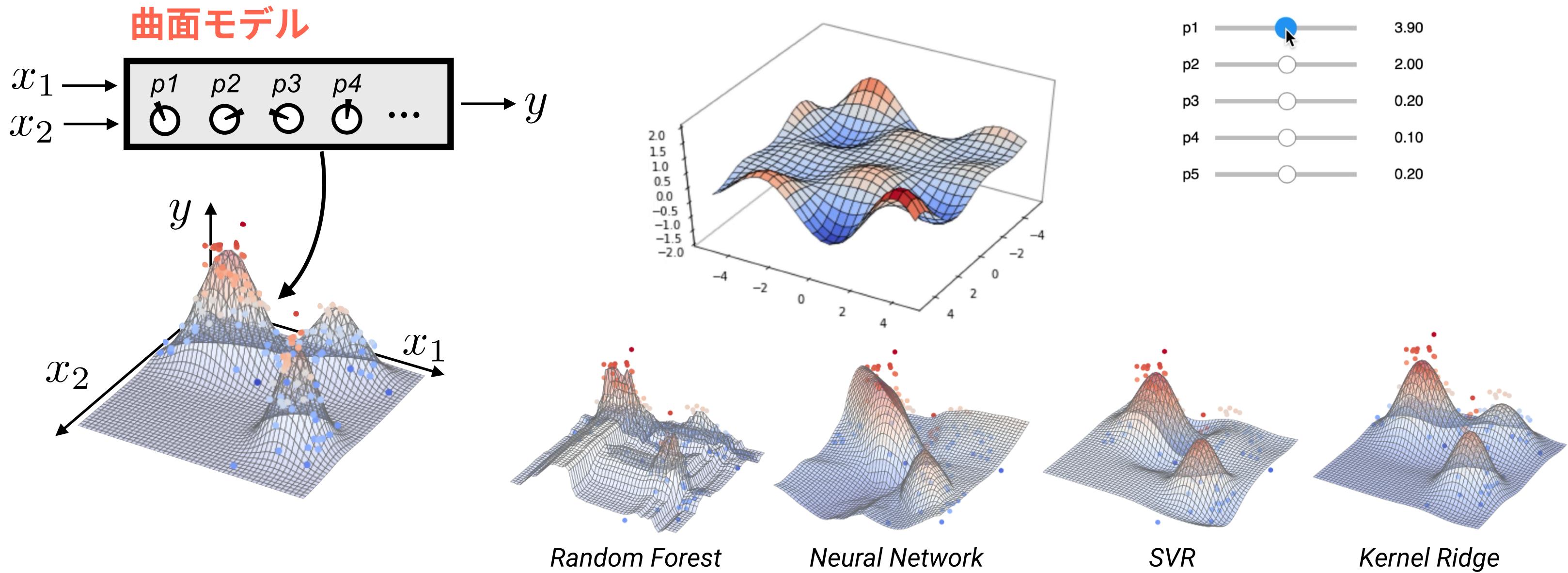
Gaussian Process



Random Forest

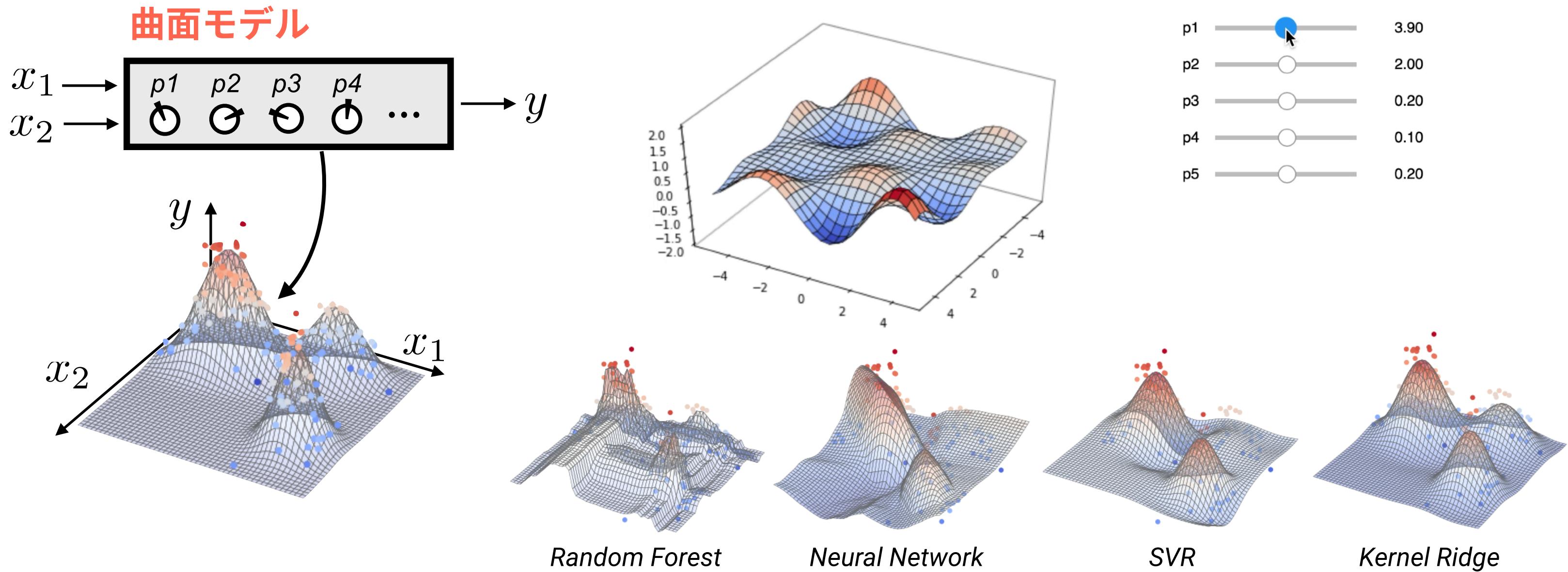
仕組みは曲面フィッティングによるデータ内挿

「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



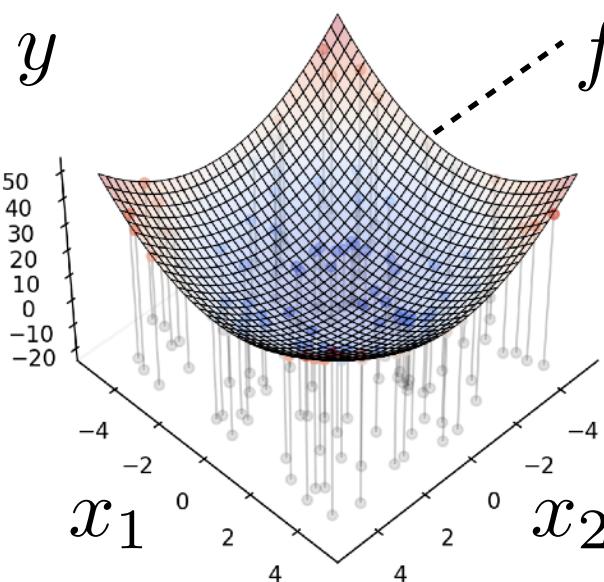
仕組みは曲面フィッティングによるデータ内挿

「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

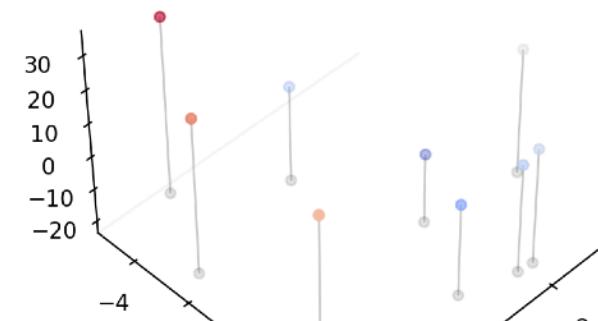


機械学習を使うのにどれくらいデータが必要ですか？

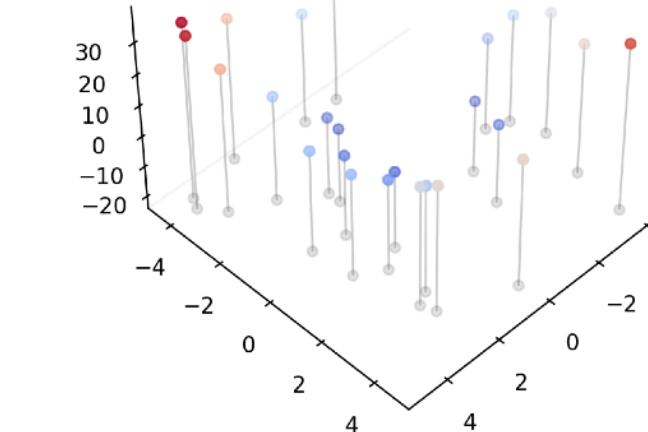
= 曲面の概形を知るのに見本点が何点必要ですか？



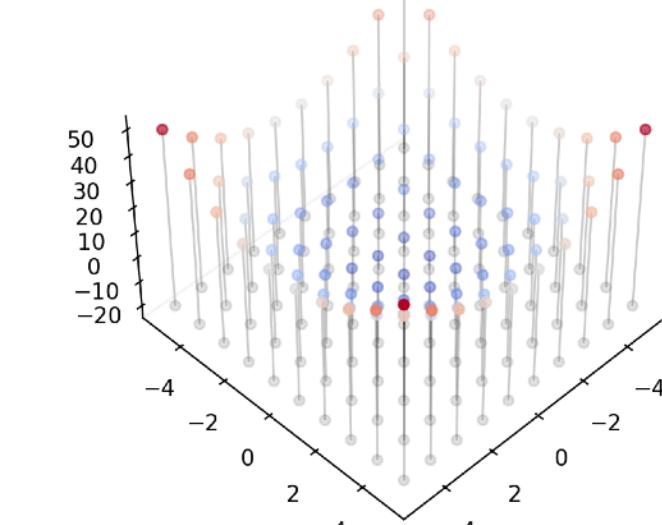
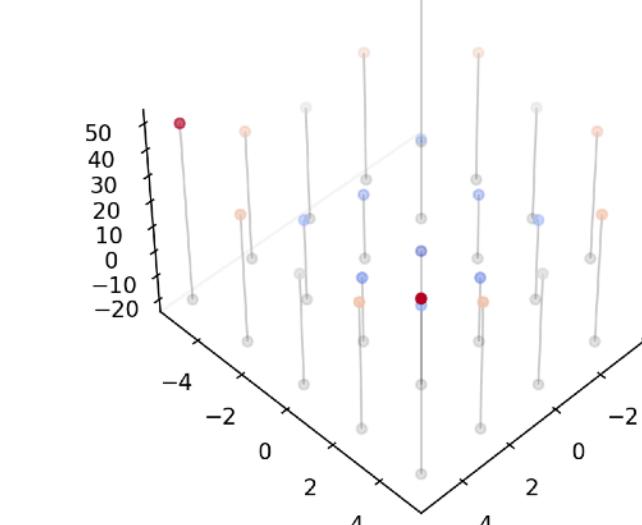
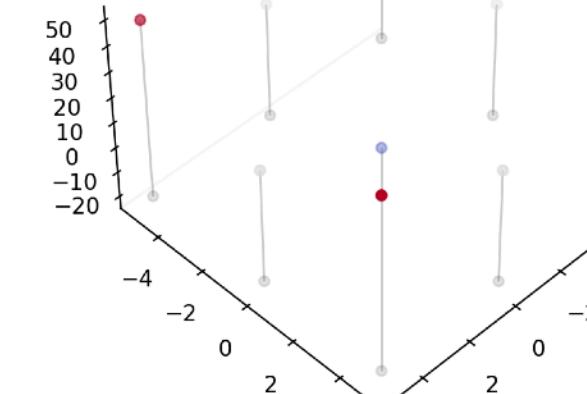
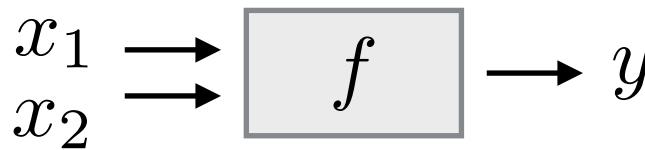
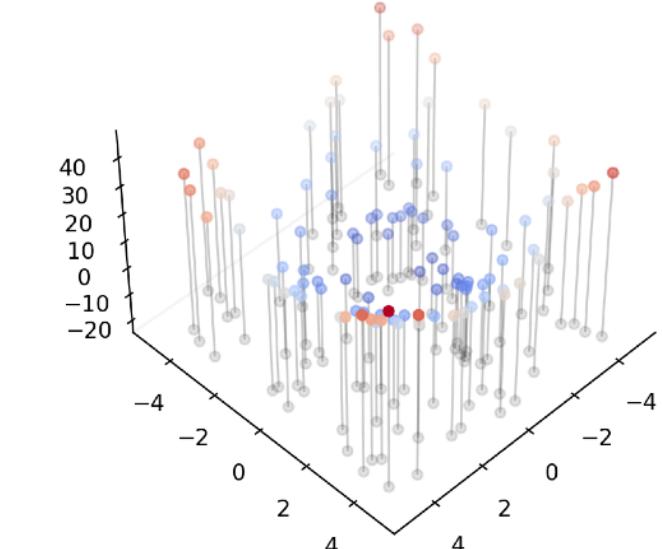
$$N = 3^2 = 9$$



$$N = 5^2 = 25$$

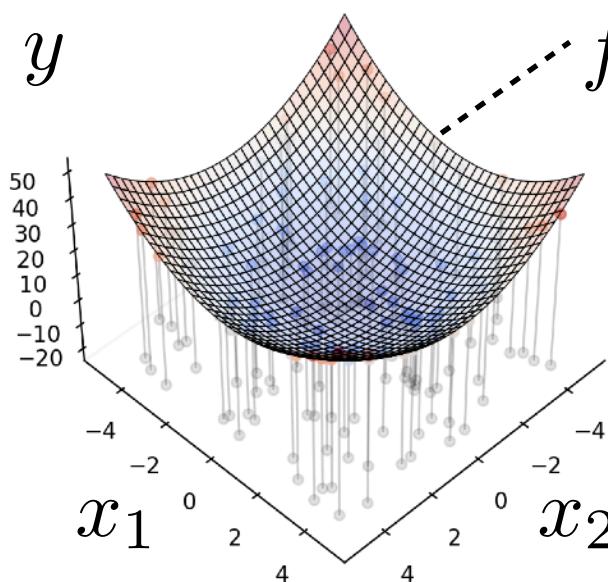


$$N = 10^2 = 100$$

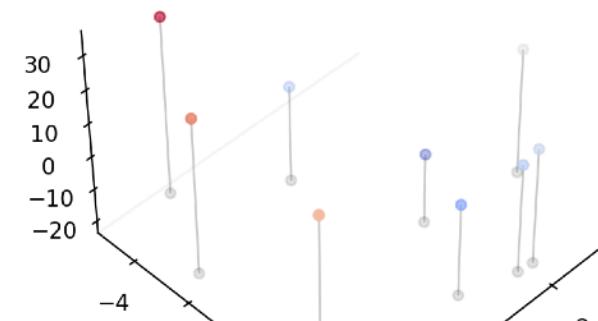


機械学習を使うのにどれくらいデータが必要ですか？

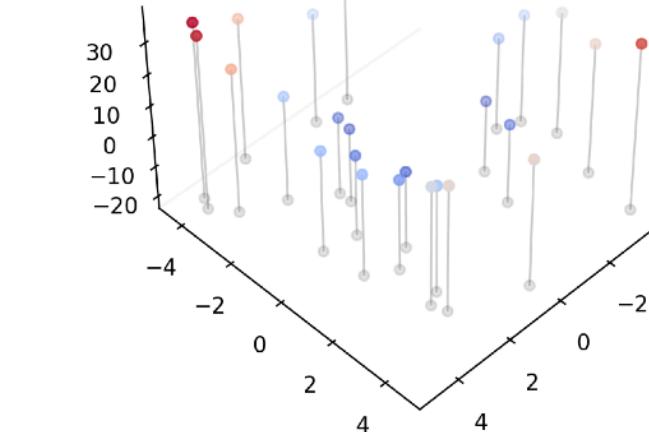
= 曲面の概形を知るために見本点が何点必要ですか？



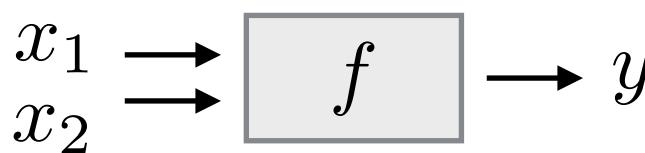
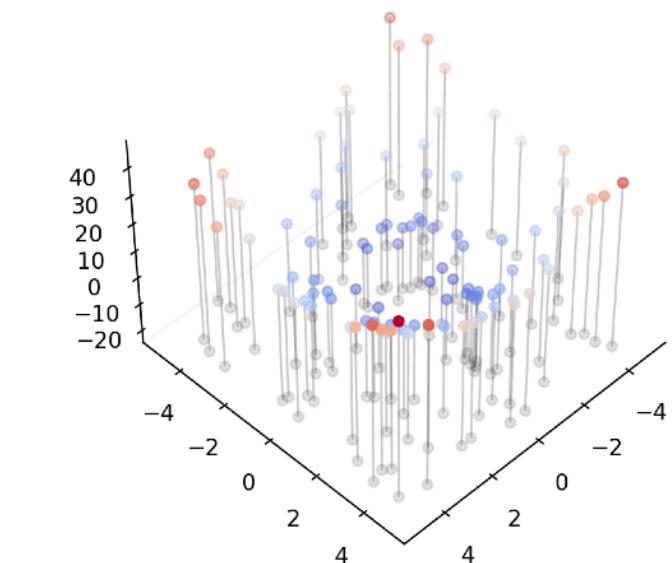
$$N = 3^2 = 9$$



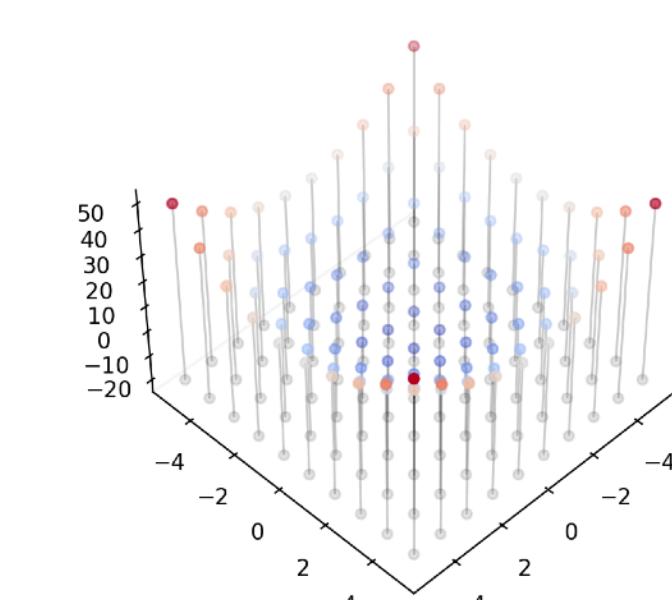
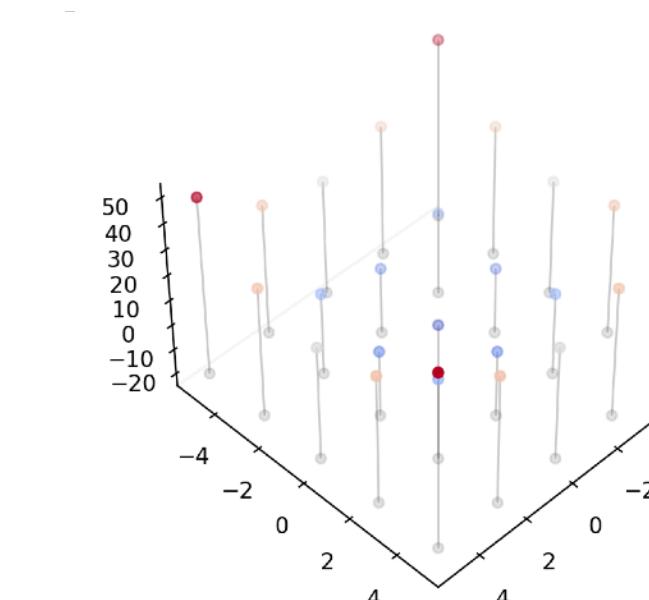
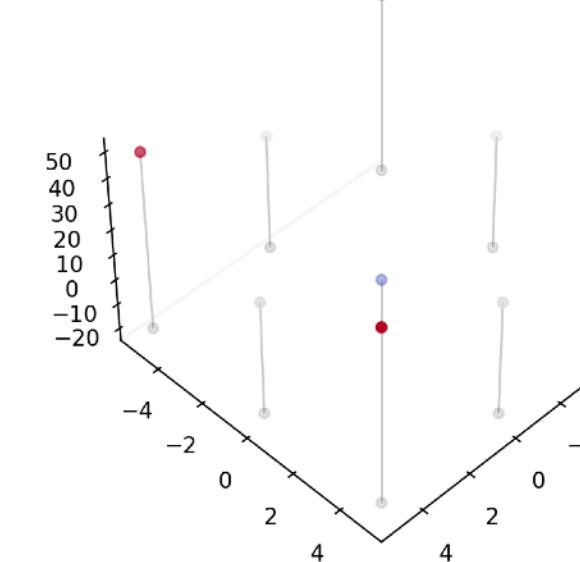
$$N = 5^2 = 25$$



$$N = 10^2 = 100$$

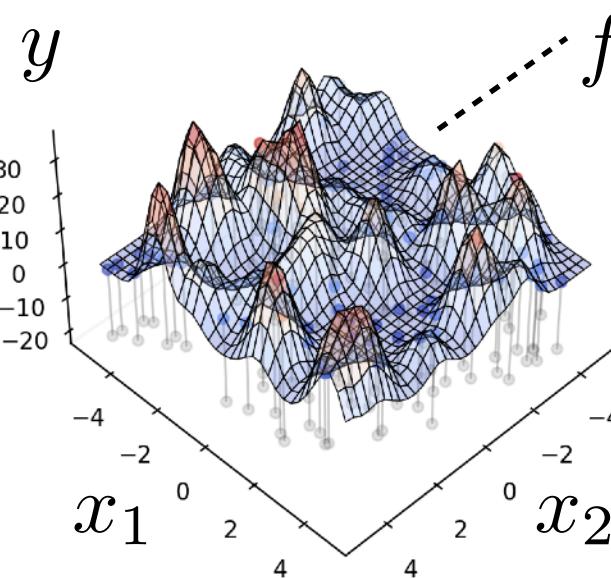


多変数(高次元)だと
データがもっと必要

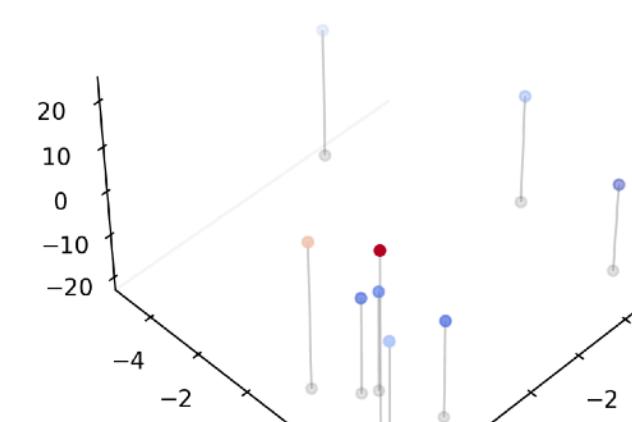


機械学習を使うのにどれくらいデータが必要ですか？

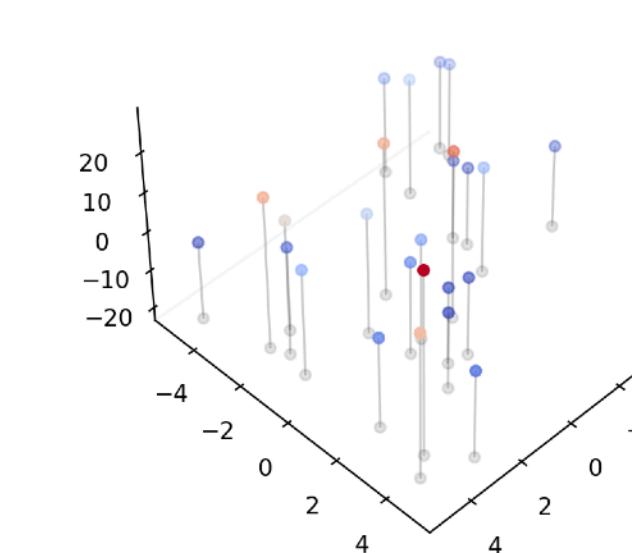
= 曲面の概形を知るのに見本点が何点必要ですか？



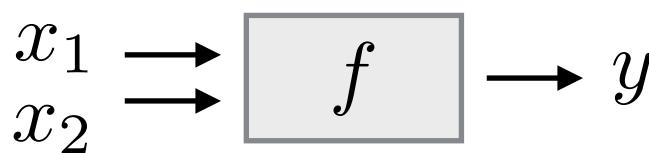
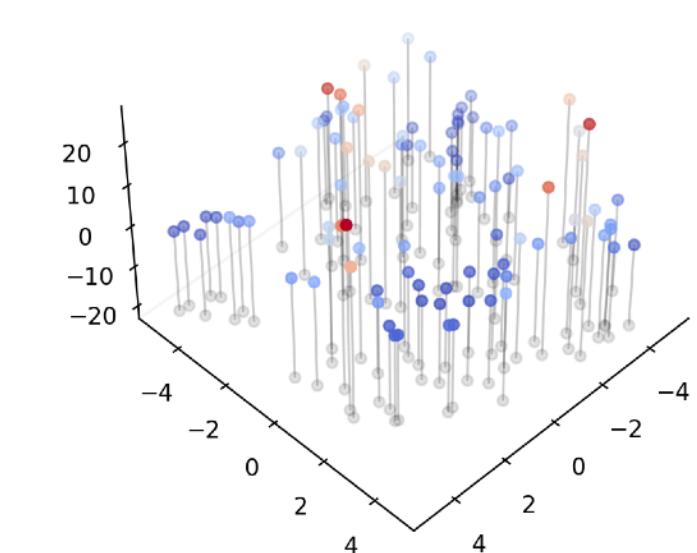
$$N = 3^2 = 9$$



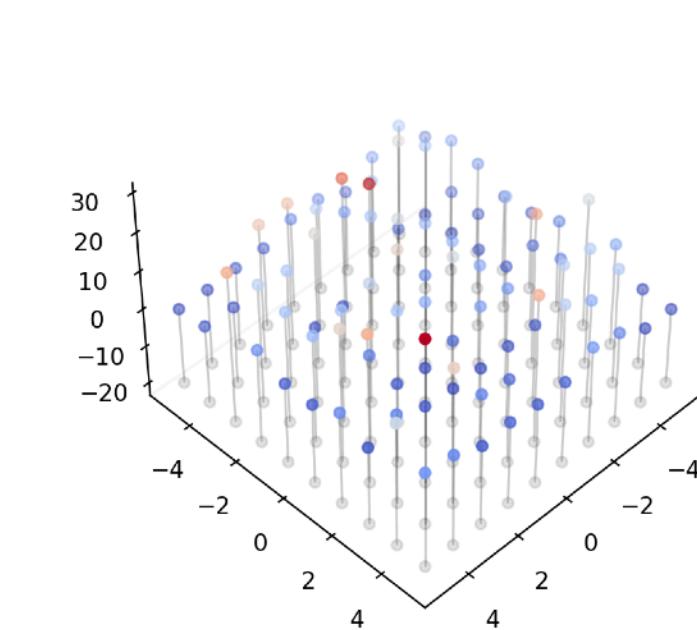
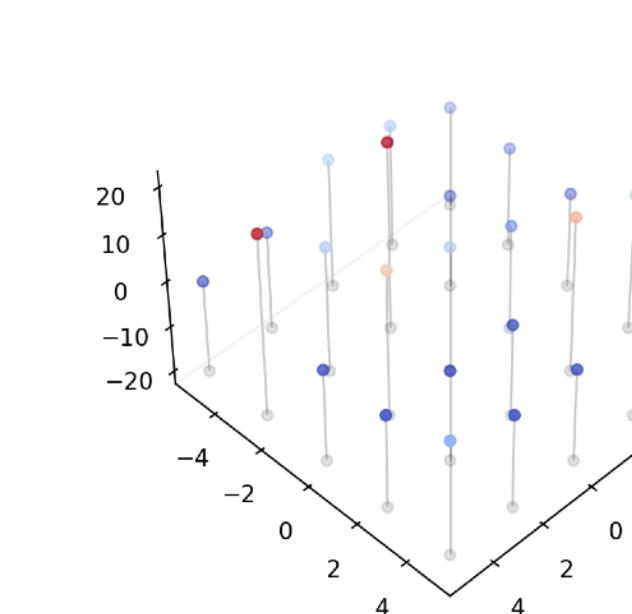
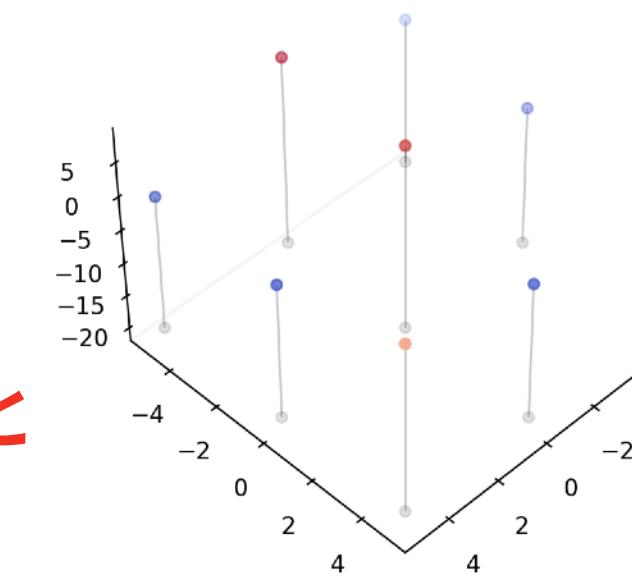
$$N = 5^2 = 25$$



$$N = 10^2 = 100$$

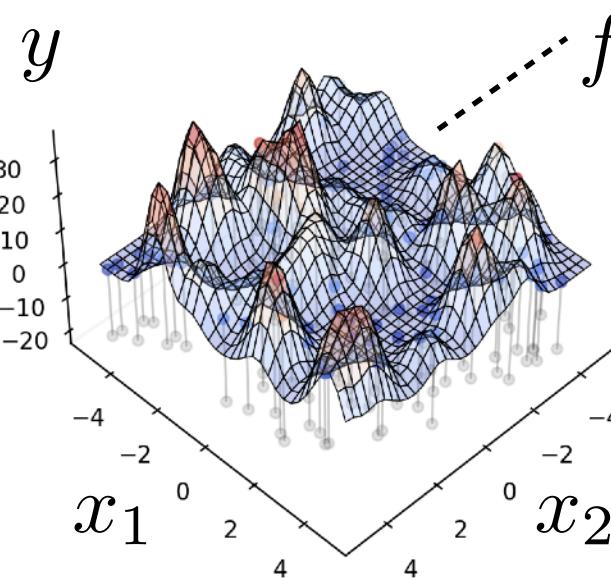


曲面(入出力)が複雑だと
もっとデータ必要

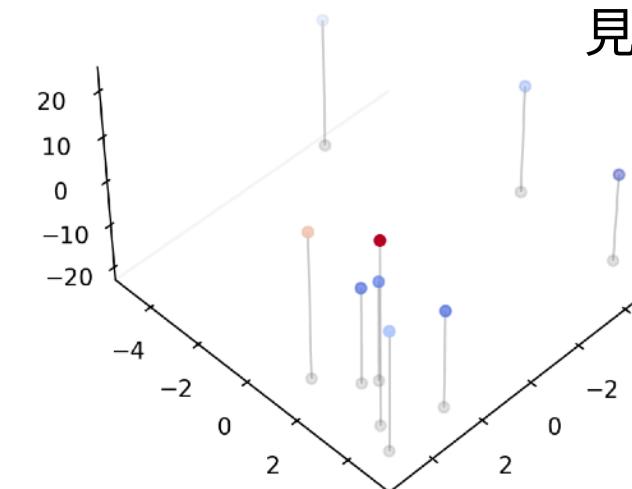


機械学習を使うのにどれくらいデータが必要ですか？

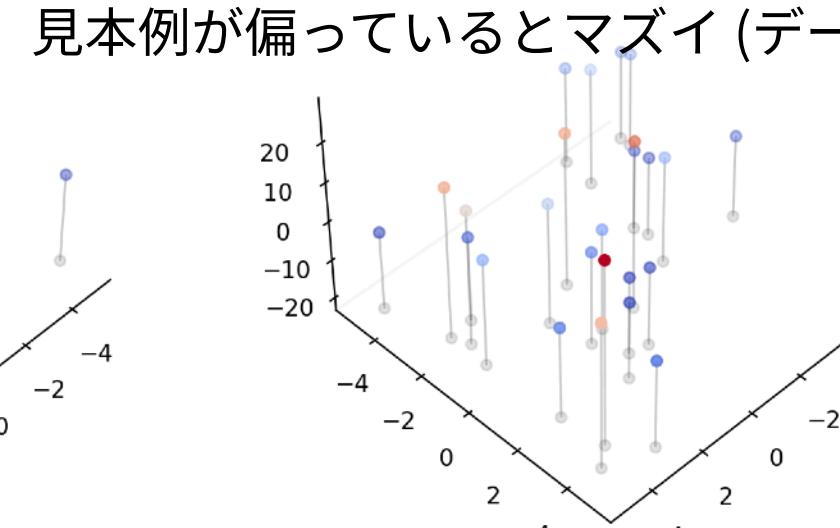
= 曲面の概形を知るのに見本点が何点必要ですか？



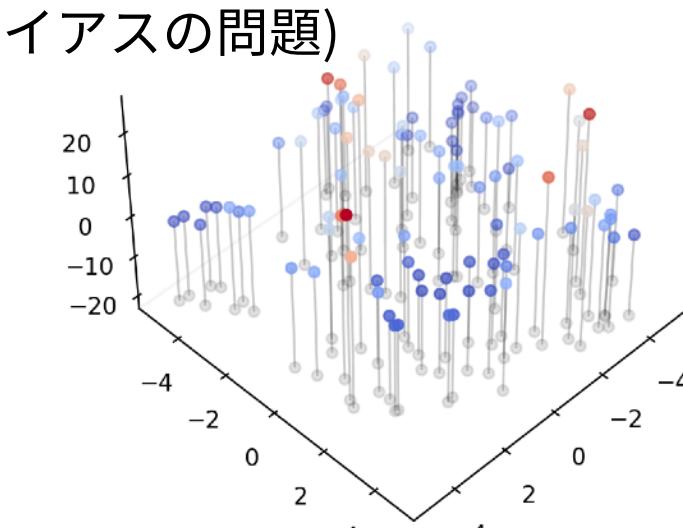
$$N = 3^2 = 9$$



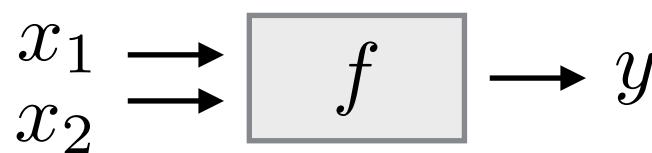
$$N = 5^2 = 25$$



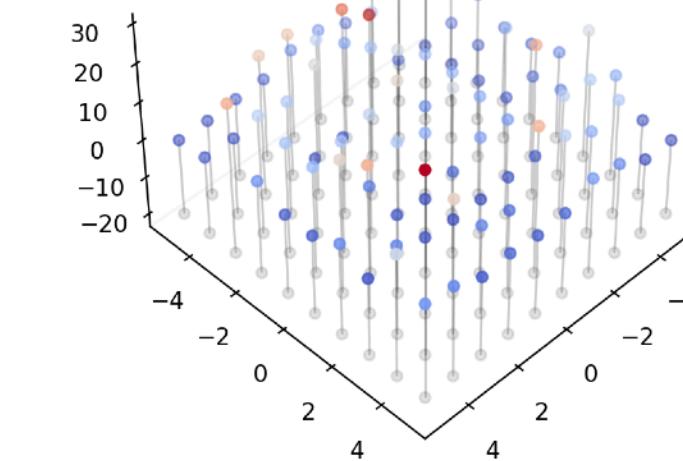
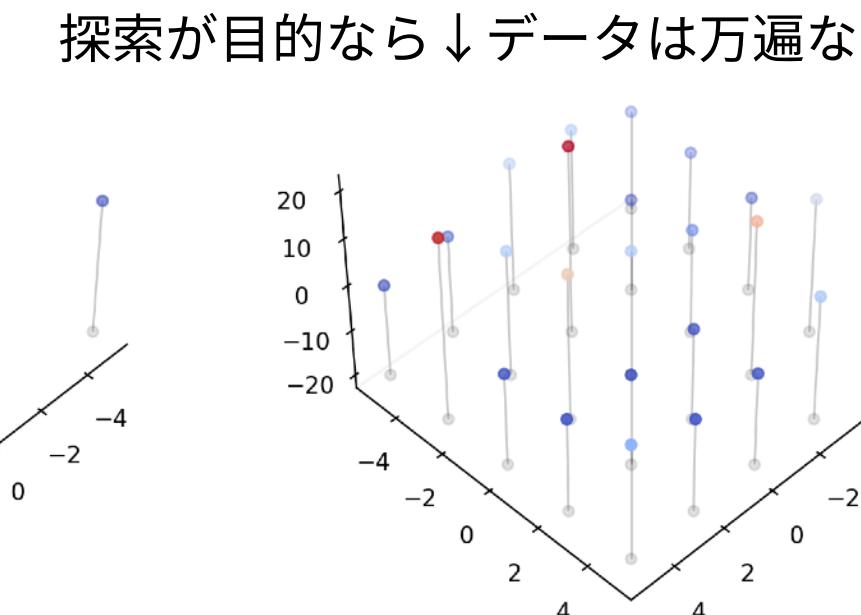
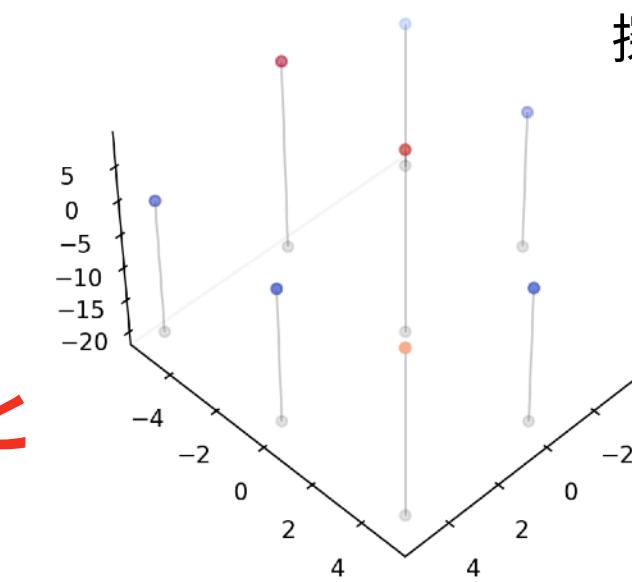
$$N = 10^2 = 100$$



見本例が偏っているとマズイ (データバイアスの問題)



曲面(入出力)が複雑だと
もっとデータ必要



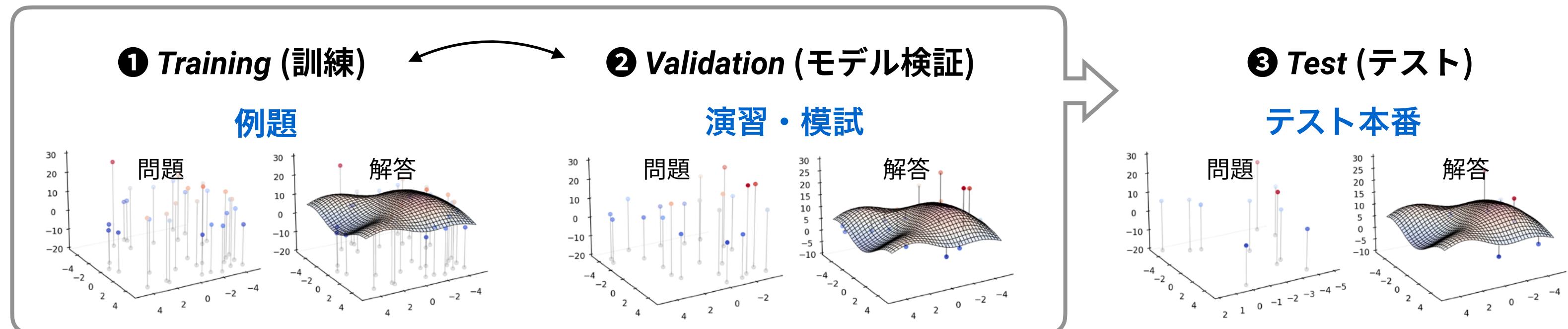
探索が目的なら↓データは万遍なく取るほうが良い (実験計画法)

機械学習の予測が当たっているかの確認には落とし穴多数！

機械学習が「与えた見本例」を正しく予測できるのは**当たり前**

真の関心：「見本例ではないデータ」に対して正しく予測できるか？

- ✓ 予測精度も手元のデータから見積もるしかないので一般にこの判断は激ムズ
- ✓ データは手元にもうあるので意図しないカンニング事故(data leakage)がとても起きやすい…



半世紀前に生まれた「希望的な呼称」による幻想にご注意

現在の機械学習は一般に想像する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」機能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…

「人工知能」「機械学習」などの希望的な呼称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

SIGART Newsletter No. 57 April 1976

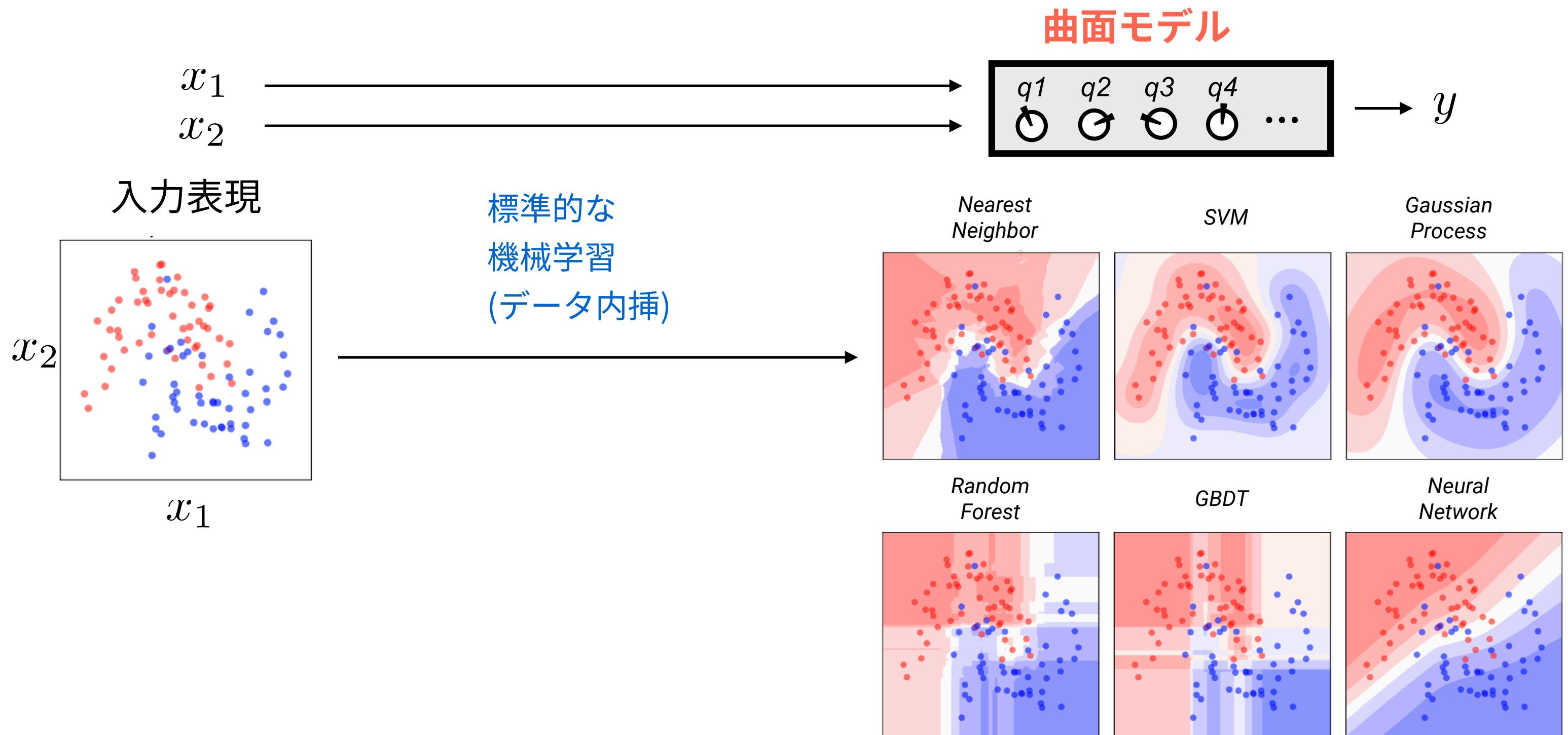
ARTIFICIAL INTELLIGENCE MEETS NATURAL STUPIDITY

Drew McDermott

MIT AI Lab Cambridge, Mass 02139

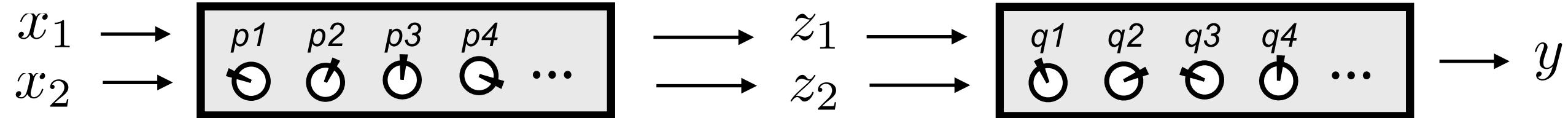
As a field, artificial intelligence has always been on the border of respectability, and therefore on the border of crackpottery. Many critics <Dreyfus, 1972>, <Lighthill, 1973> have urged that we are over the border. We have been very defensive toward this charge, drawing ourselves up with dignity when 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

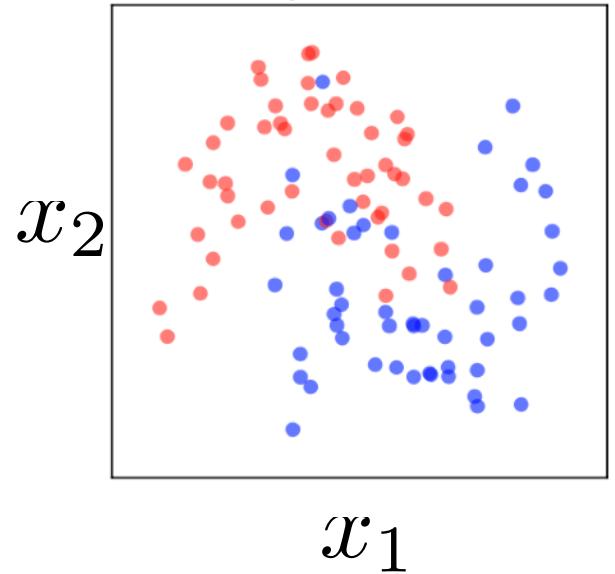


深層学習と表現学習

変数変換(表現学習)

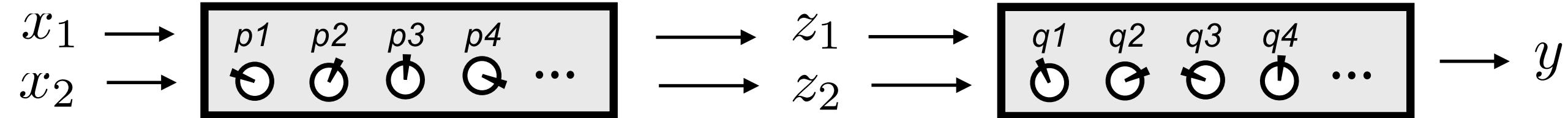


入力表現



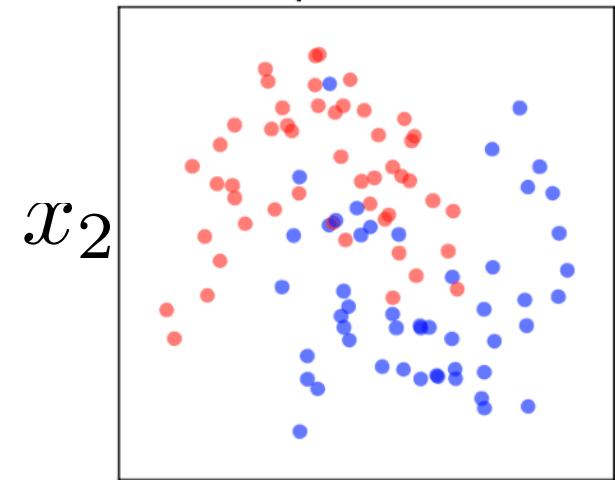
深層学習と表現学習

変数変換(表現学習)

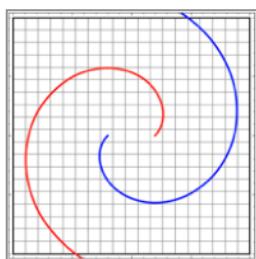


曲面モデル

入力表現

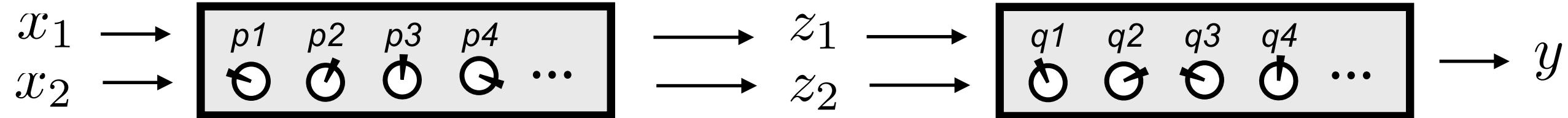


x_1



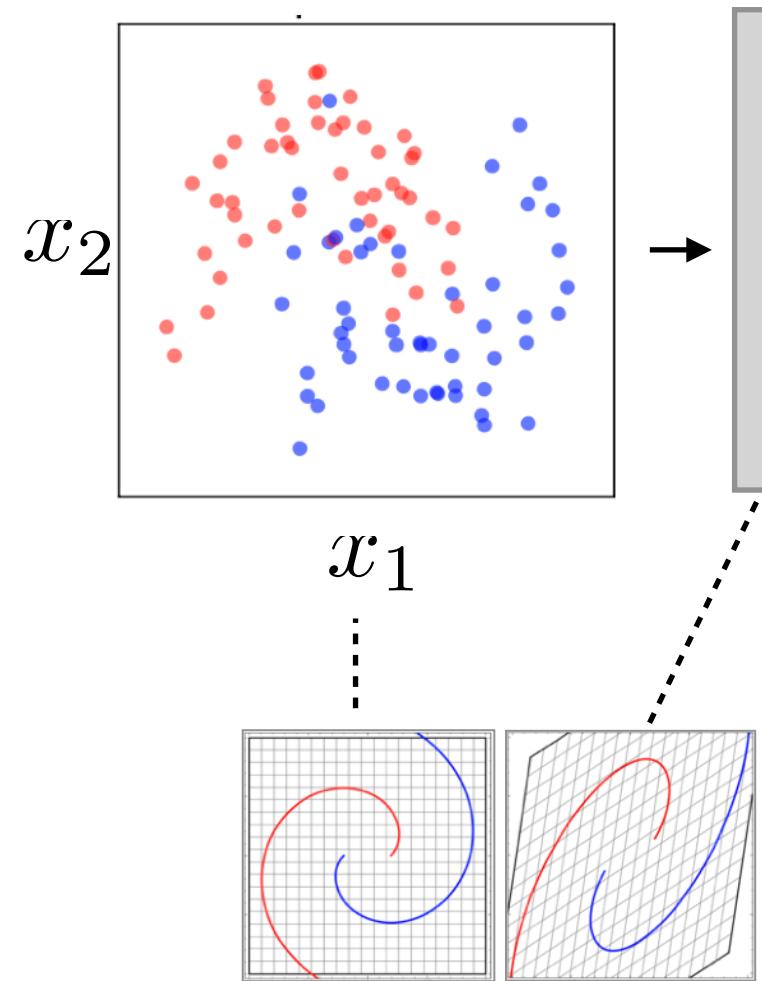
深層学習と表現学習

変数変換(表現学習)



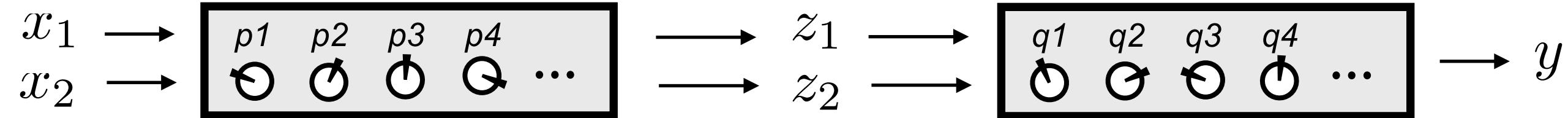
曲面モデル

入力表現



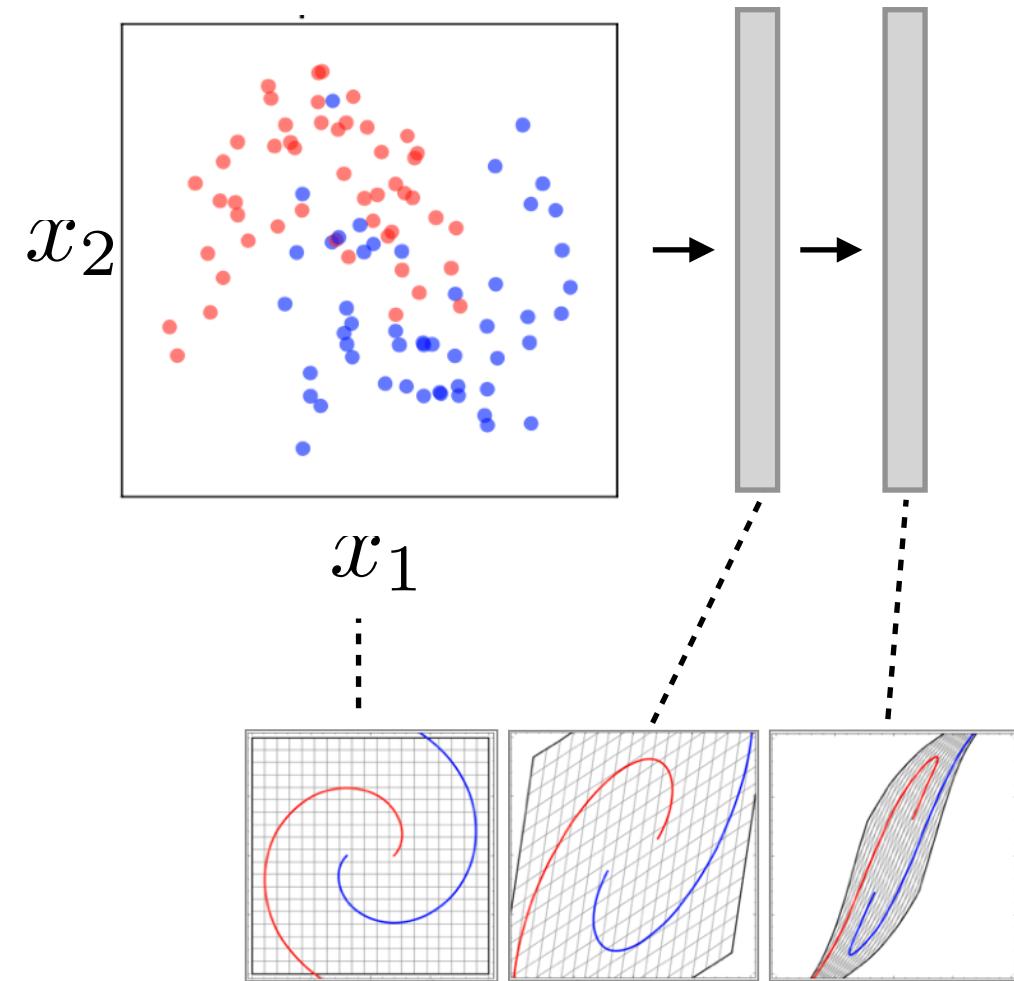
深層学習と表現学習

変数変換(表現学習)



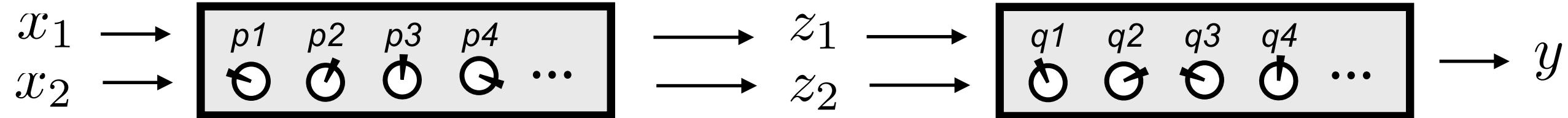
曲面モデル

入力表現



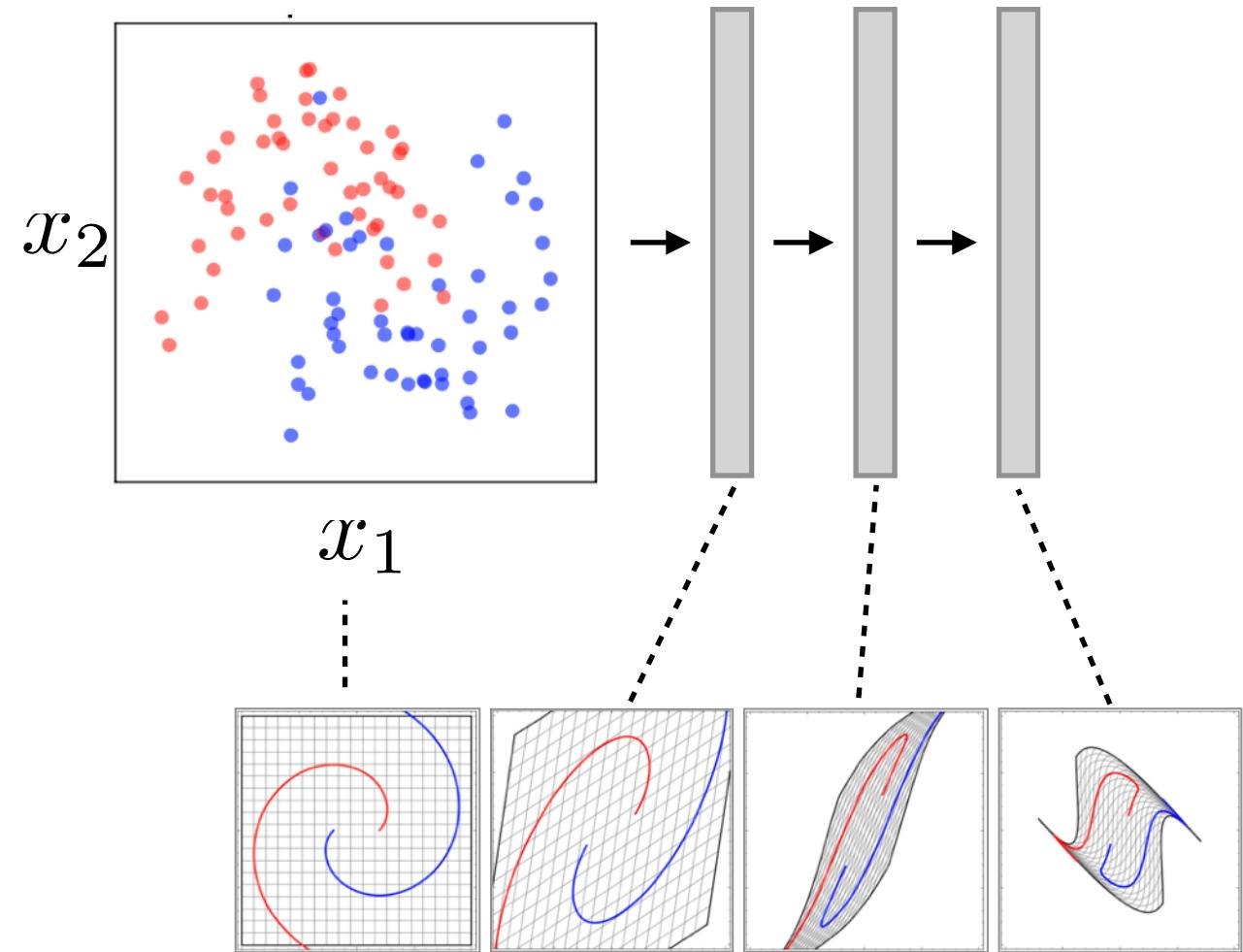
深層学習と表現学習

変数変換(表現学習)



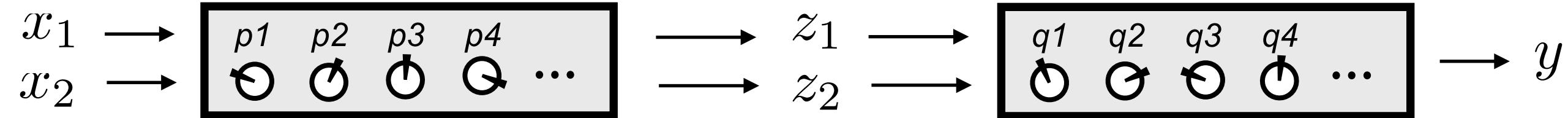
曲面モデル

入力表現

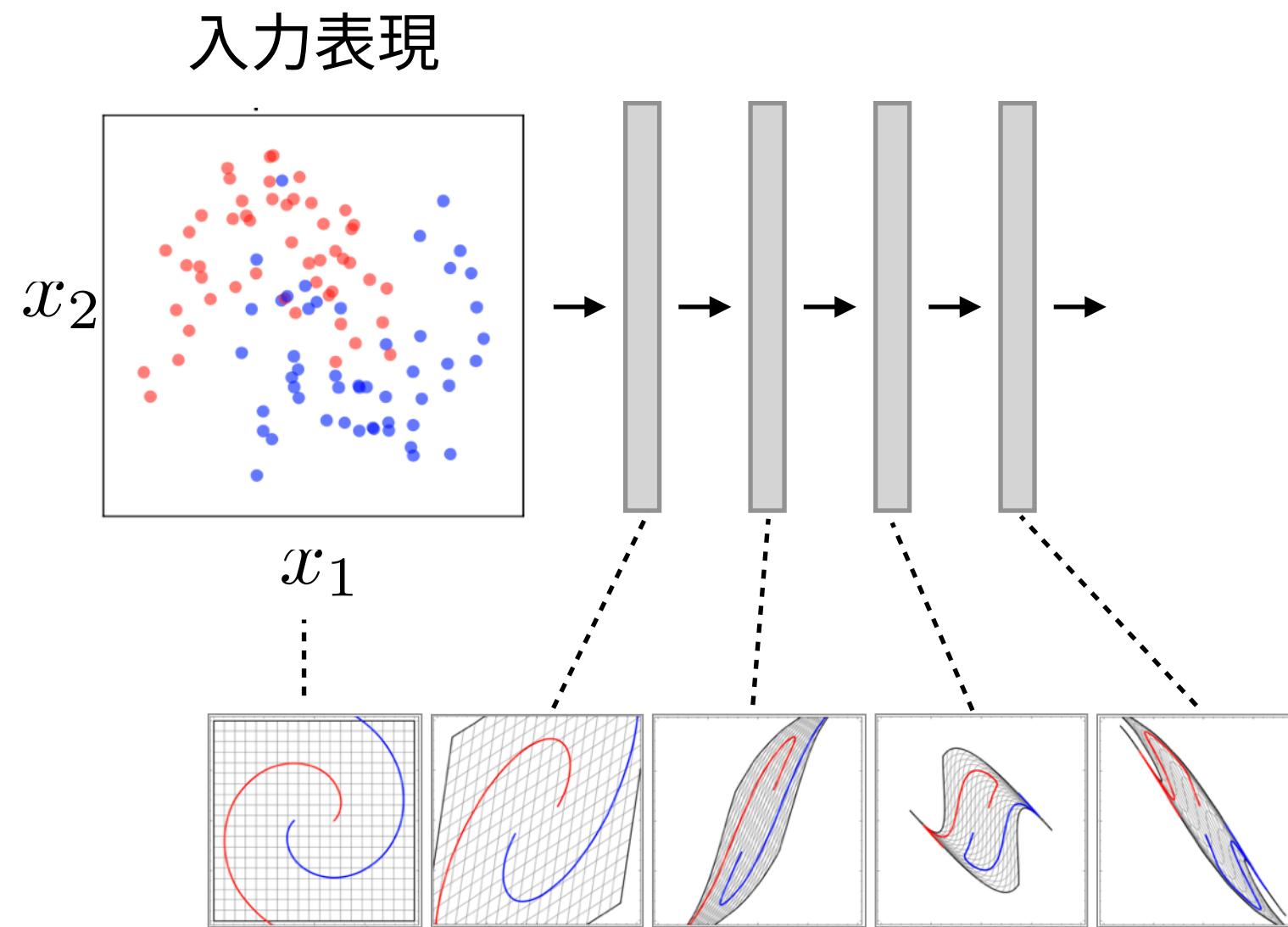


深層学習と表現学習

変数変換(表現学習)

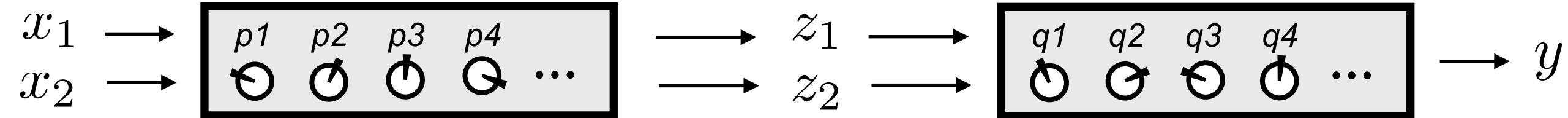


曲面モデル

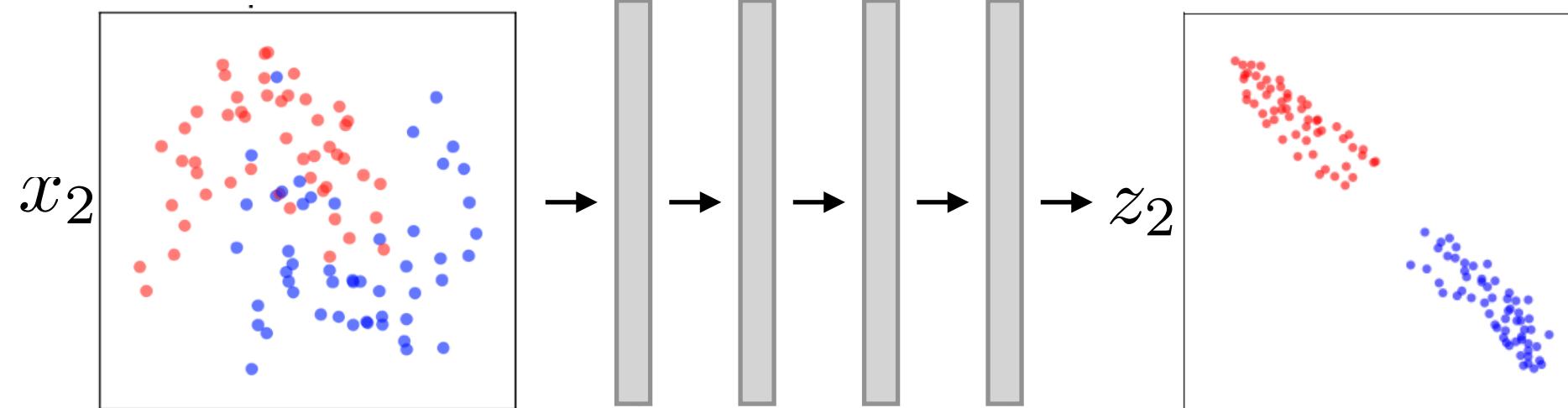


深層学習と表現学習

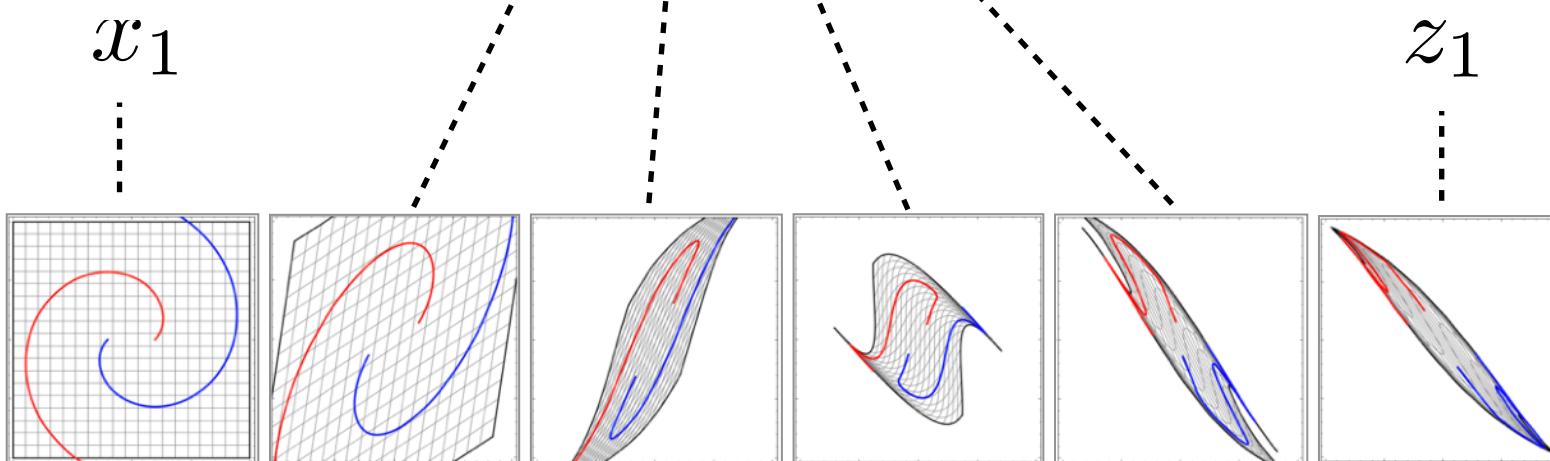
変数変換(表現学習)



入力表現



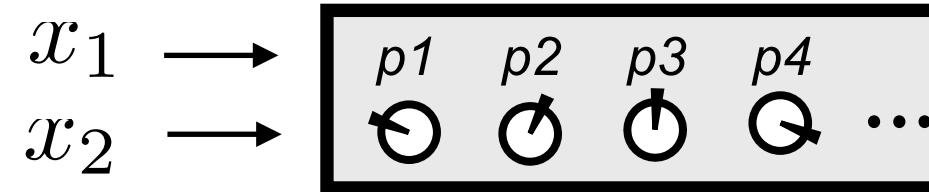
良い表現



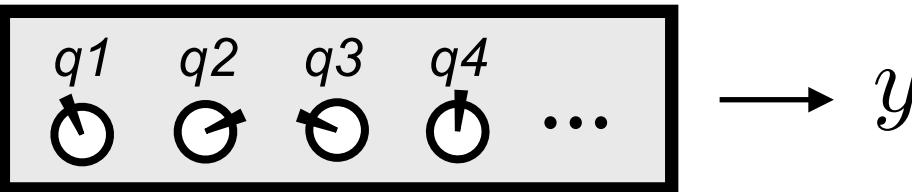
曲面モデル

深層学習と表現学習

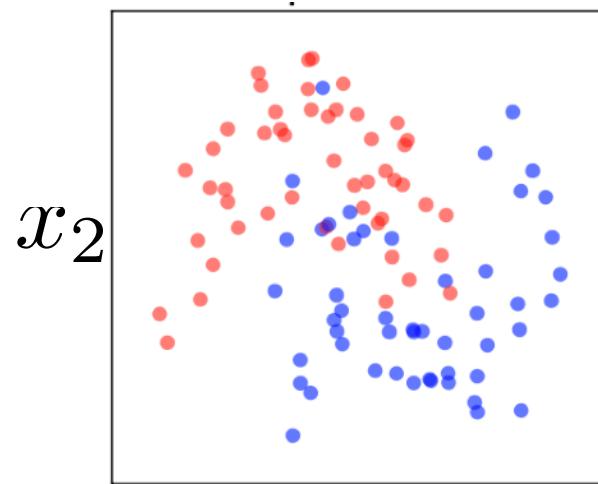
変数変換(表現学習)



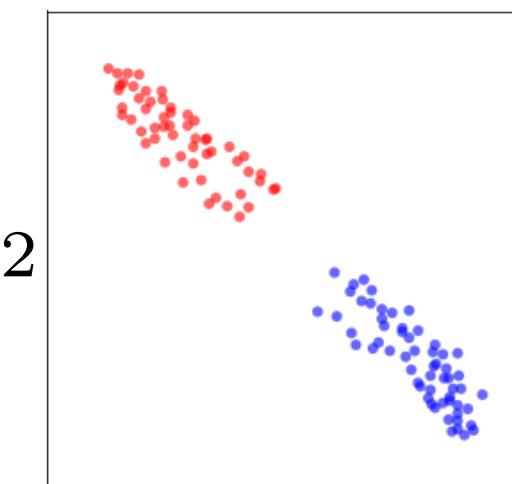
曲面モデル



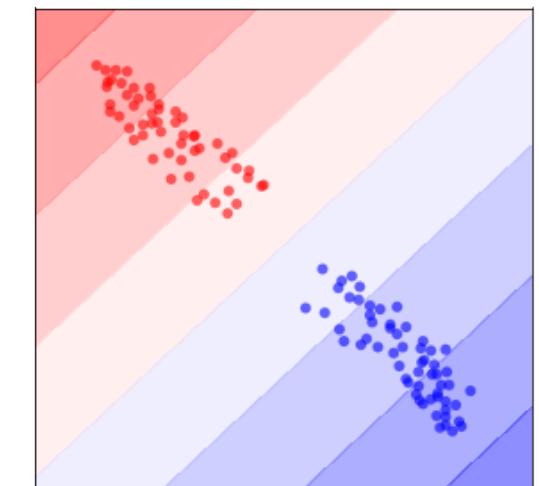
入力表現



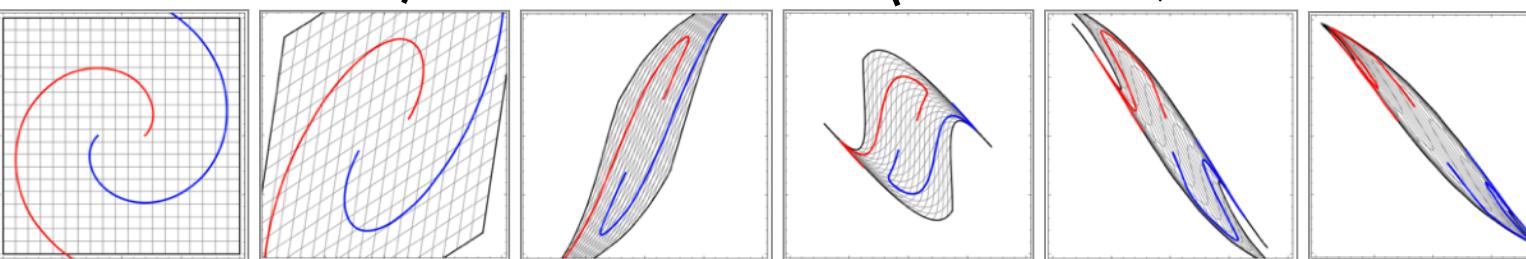
良い表現



標準的な
機械学習
(データ内挿)



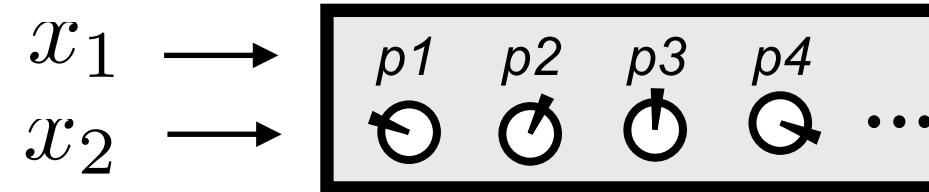
x_1



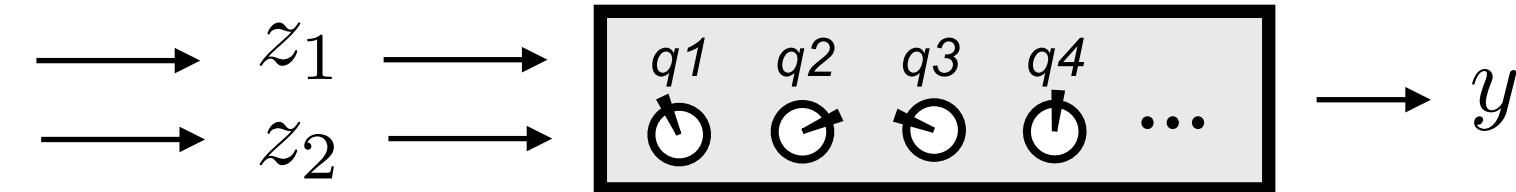
シンプル(線形)で十分

深層学習と表現学習

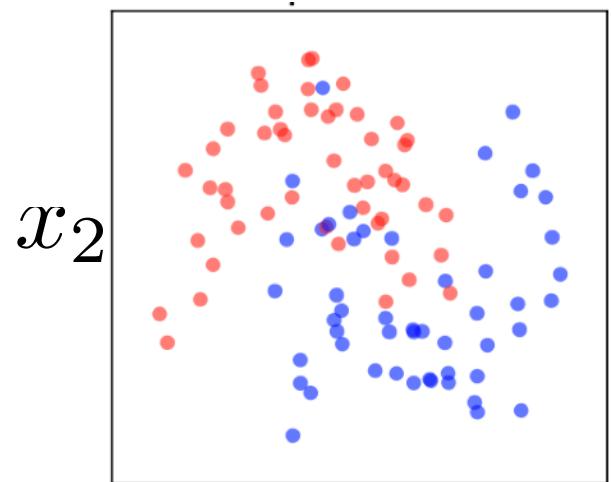
変数変換(表現学習)



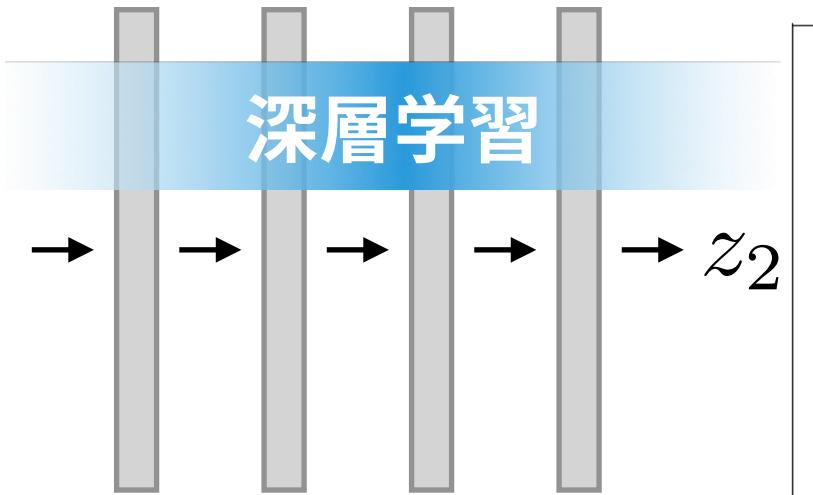
曲面モデル



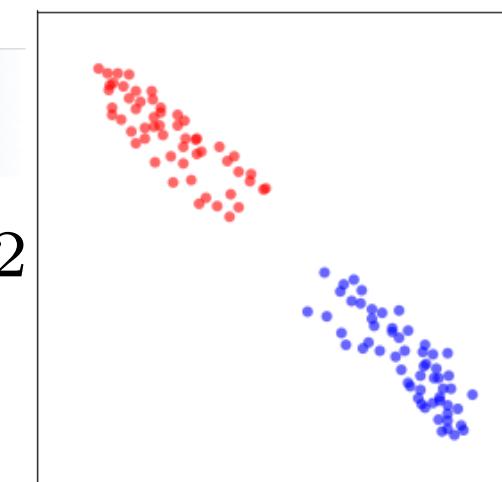
入力表現



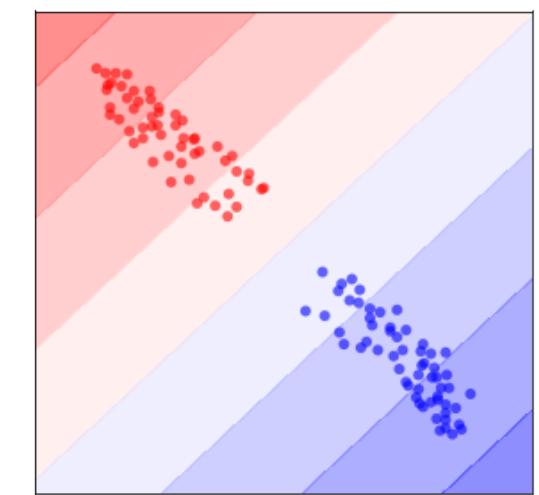
深層学習



良い表現



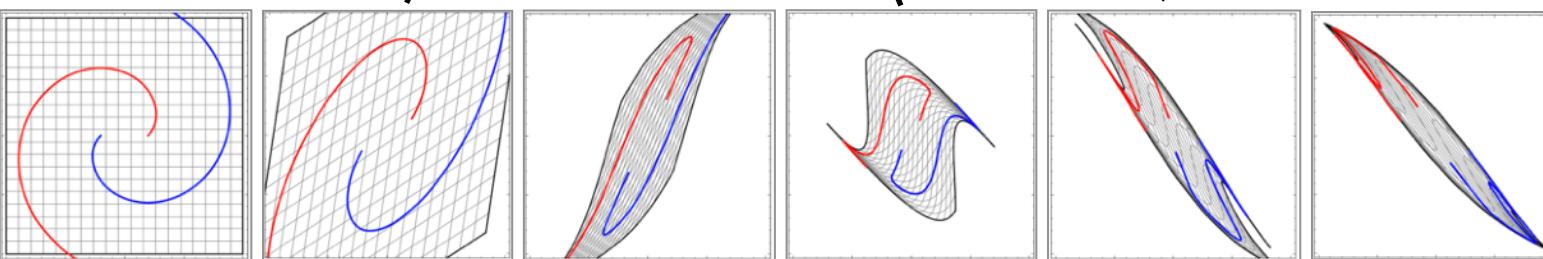
標準的な
機械学習
(データ内挿)



x_1

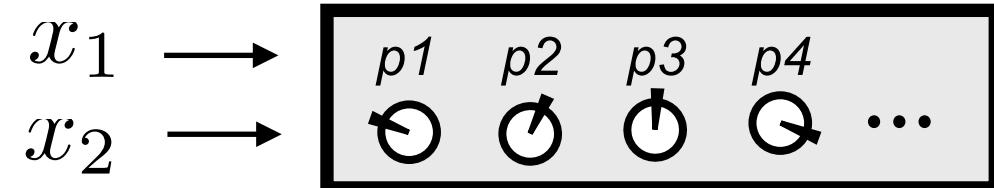
z_1

シンプル(線形)で十分

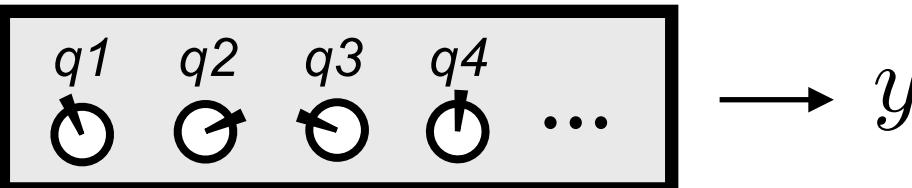


深層学習と表現学習

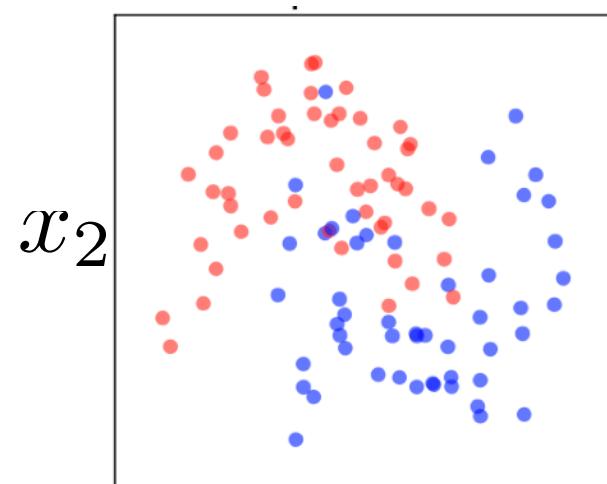
変数変換(表現学習)



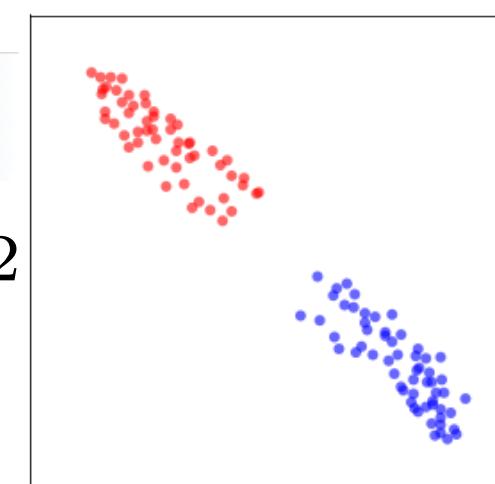
曲面モデル



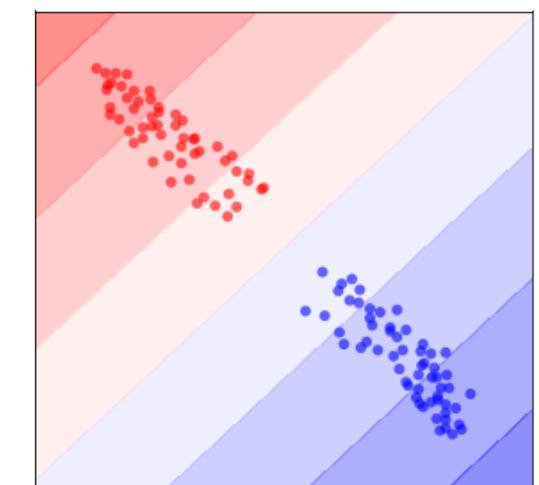
入力表現



良い表現

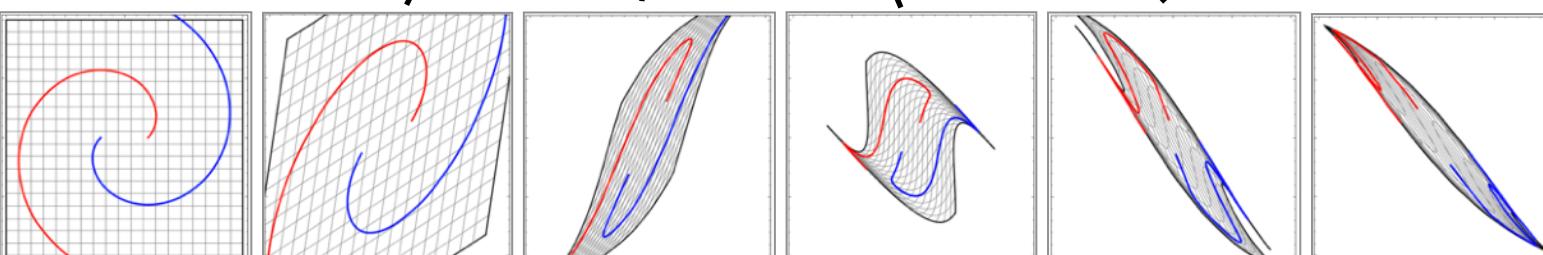


標準的な
機械学習
(データ内挿)



x_1

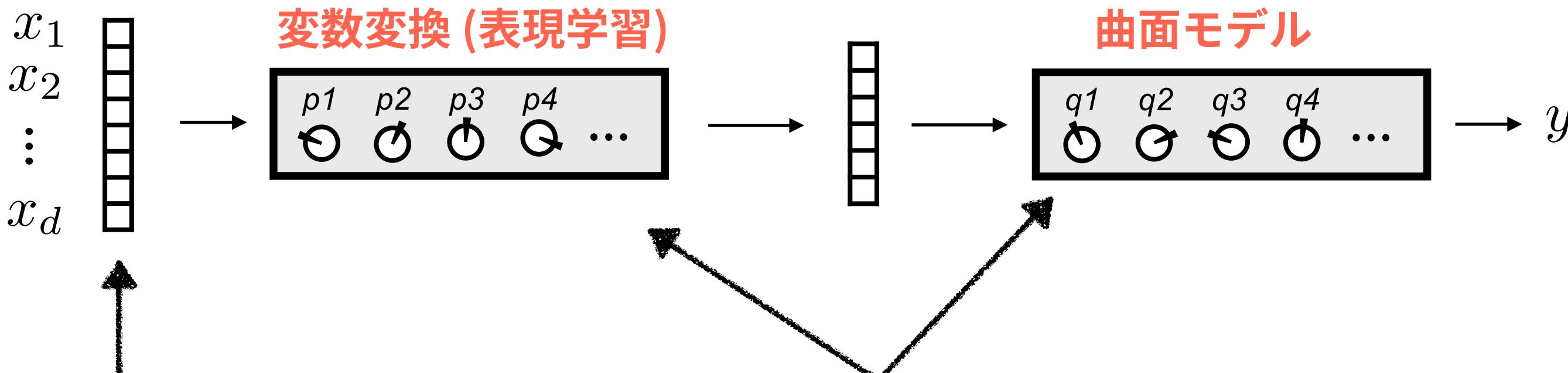
z_1



シンプル(線形)で十分

データ内挿は入力表現ではなく
「良い表現」で行う

困難① 現代の機械学習モデルは多量のデータを必要とする



① 高次元性：入力変数が多すぎ！

- ✓ 機械学習は入力されてない情報を全く考慮してくれない… (擬似相関リスク)
- ✓ とりあえず色々な変数を入れがち

画像そのままを入力する場合

20×20 ピクセルのカラー画像 → 1200変数

1000×1000 ピクセルのカラー画像 → 300万変数

② 過剰パラメタ化：パラメタ数が多すぎ！

画像 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

言語 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

困難② 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）

高い予測精度を持つ機械学習モデルは一つのデータセットからたくさん作れる！

有限データから見積もる予測精度では指数的広さのモデル探索空間をspecifyしきれない…

困難② 羅生門効果とUnderspecification

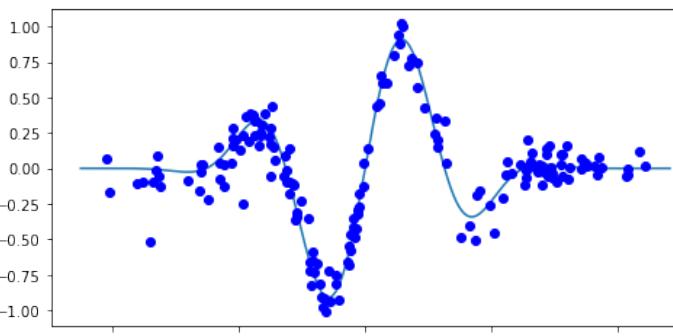
羅生門効果：良い機械学習モデルの多重性（非一意性）

高い予測精度を持つ機械学習モデルは一つのデータセットからたくさん作れる！

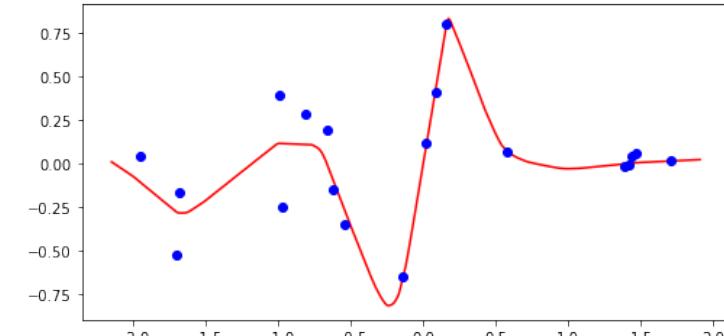
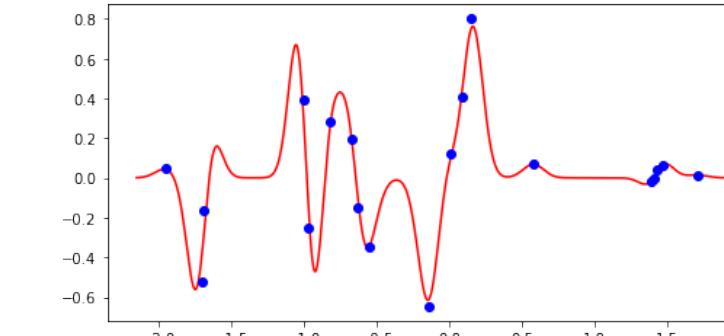
有限データから見積もる予測精度では指数的広さのモデル探索空間をspecifyしきれない…

- ✓ 「どのモデルが求める真実なの？」と考えてしまうと、まさに真実は「**藪の中**」…
→ **複数の手法による多角的解釈**が鉄則 "all models are wrong but some models are useful"
- ✓ 実際には**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



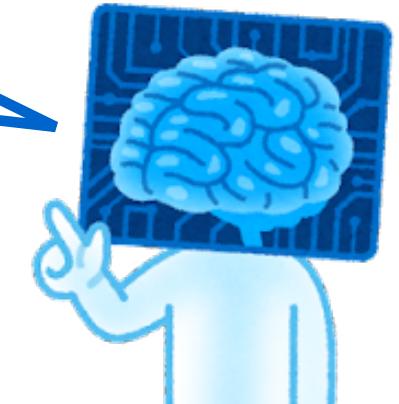
手法やモデルによって予測時の挙動にかなり差が出てしまう



Underspecificationはビッグデータ事例でも広く起こっている？

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



[https://ai.googleblog.com/2021/10/
how-underspecification-presents.html](https://ai.googleblog.com/2021/10/how-underspecification-presents.html)



The latest from Google Research

How Underspecification Presents Challenges for Machine
Learning

Monday, October 18, 2021

Posted by Alex D'Amour and Katherine Heller, Research Scientists, Google Research

Machine learning (ML) models are being used more widely today than ever before and are becoming increasingly impactful. However, they often exhibit unexpected behavior when they are used in real-world domains. For example, computer vision models can exhibit surprising sensitivity to irrelevant features, while natural language processing models can depend unpredictably on demographic correlations not directly indicated by the text. Some reasons for these failures are well-known: for example, training ML models on poorly curated data, or training models to solve prediction problems that are structurally mismatched with the application domain. Yet, even when

<https://arxiv.org/abs/2011.03395>

arXiv.org > cs > arXiv:2011.03395

Search...

Help | Advanced S

Computer Science > Machine Learning

[Submitted on 6 Nov 2020 (v1), last revised 24 Nov 2020 (this version, v2)]

Underspecification Presents Challenges for Credibility in Modern Machine Learning

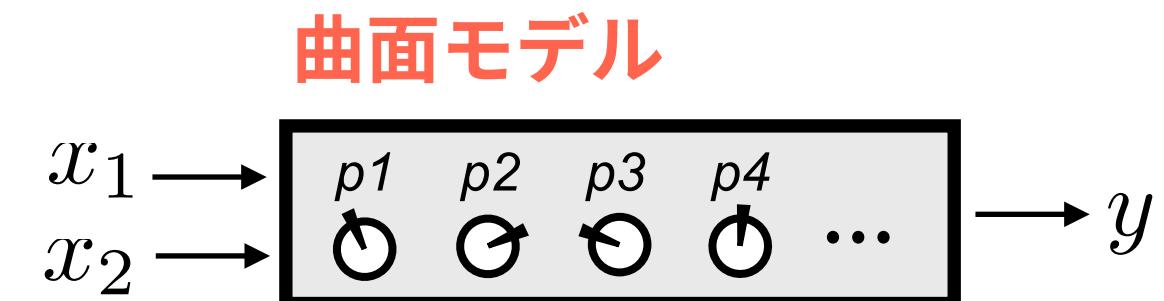
Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, D. Sculley

ML models often exhibit unexpectedly poor behavior when they are deployed in real-world domains. We identify underspecification as a key reason for these failures. An ML pipeline is underspecified when it can return many predictors with equivalently strong held-out performance in the training domain. Underspecification is common in modern ML pipelines, such as those based on deep learning. Predictors

現代の技術的関心はこの高次元性をどう手懐けるか

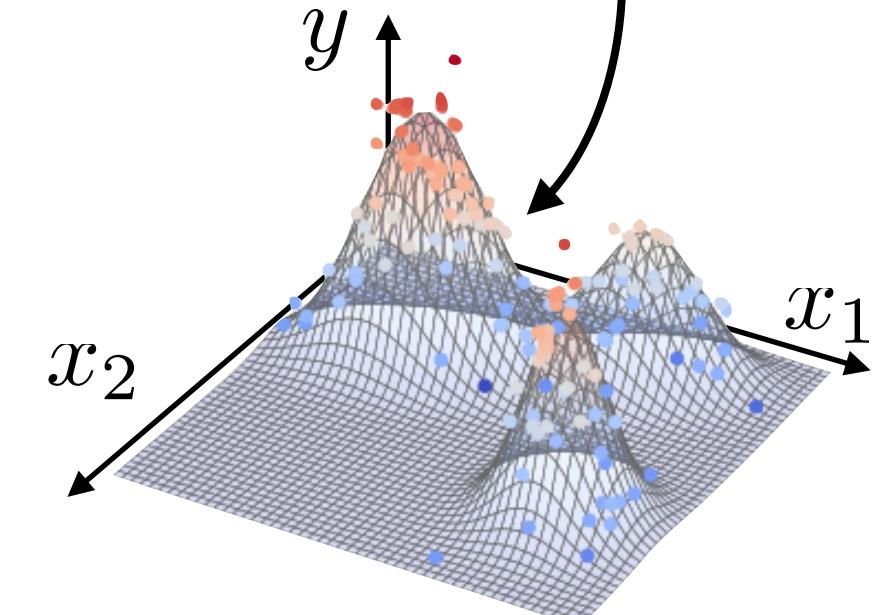
1. 確率的最適化・正則化 → モデルが大きい自由度の中で暴れまくらないよう動ける範囲を何とかして制御・制限・安定化する
2. 事前学習 (Warm Start) の転移 → 事前に得ておいたイイ感じのパラメタ初期値を使う
3. 帰納バイアスの設計

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



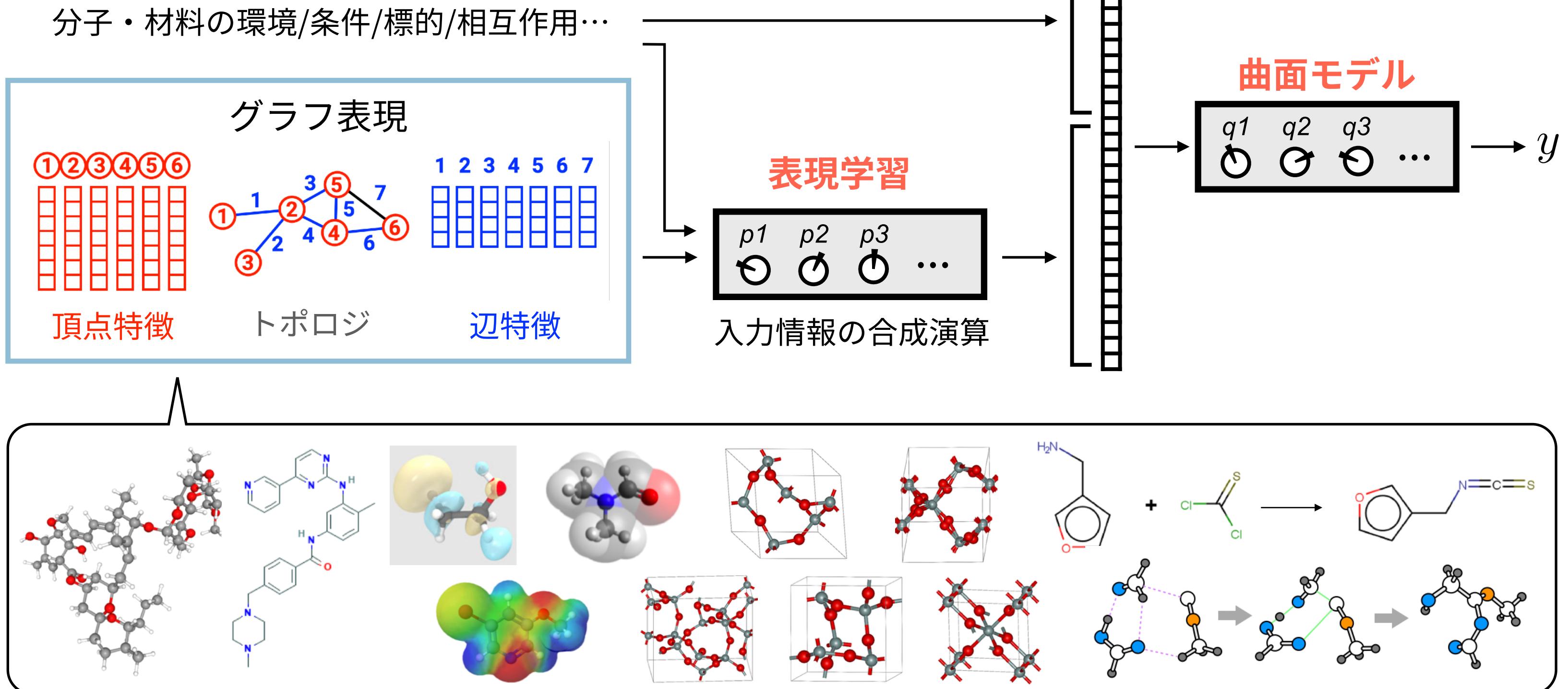
機械学習×化学：化学に適合した帰納バイアスのデザイン

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！



実例：分子の表現学習とGraph Neural Networks (GNNs)

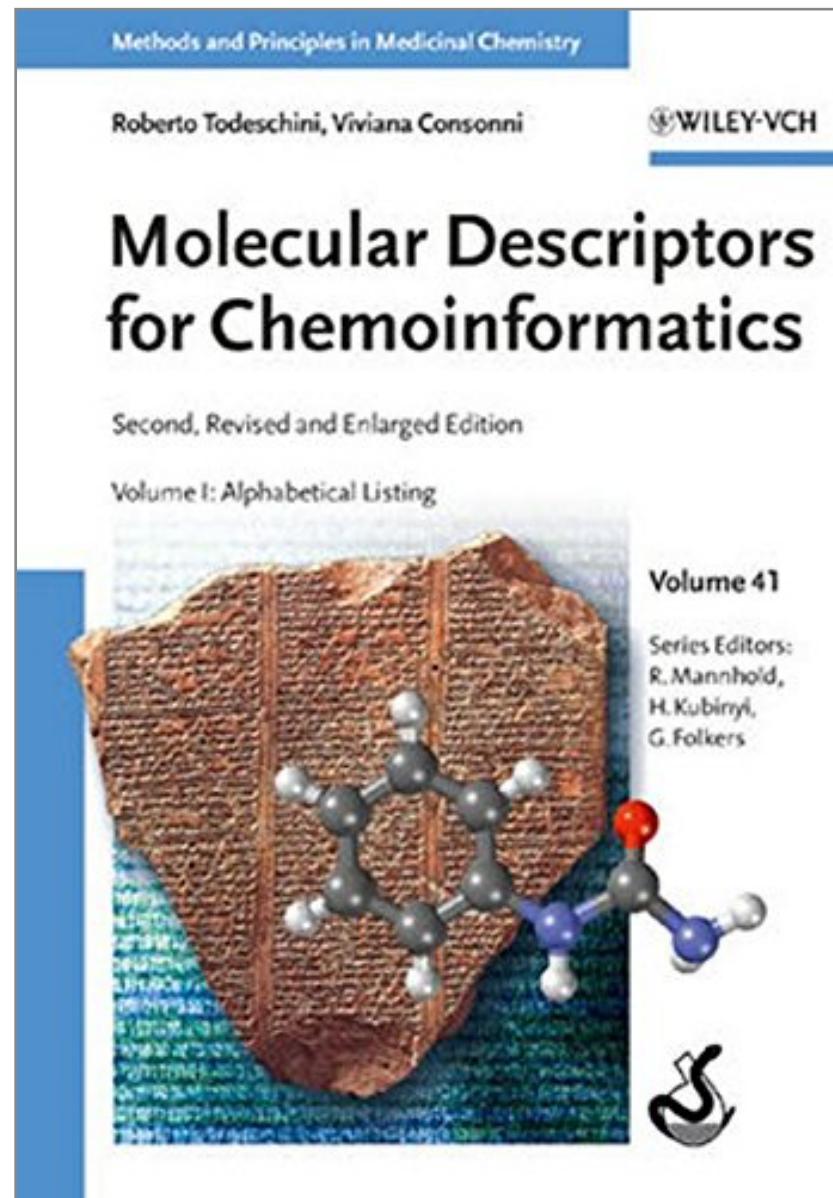
分子・材料の環境/条件/標的/相互作用…



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？

- **0-Dimensional Descriptors**
 - Constitutional descriptors
 - Count descriptors
- **1-Dimensional Descriptors**
 - List of structural fragments
 - Fingerprints
- **2-Dimensional Descriptors**
 - Graph invariants
- **3-Dimensional Descriptors**
 - 3D MoRSE, WHIM, GETAWAY, ...
 - Quantum-chemical descriptors
 - Size, steric, surface, volume, ...
- **4-Dimensional Descriptors**
 - GRID, CoMFA, Volsurf, ...

"Vol 1 contains an alphabetical listing of **more than 3,300 descriptors**"



"Dragon calculates **5,270 molecular descriptors**"

BLOCK NO	BLOCK NAME	DESCRIPTORS
1	Constitutional	47
2	Ring descriptors	32
3	Topological indices	75
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	50
7	2D matrix-based descriptors	607
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P-VSA-like descriptors	55

Use Case ① 定量的活性/物性相関 (QSAR/QSPR)



ML

ノイズが多く複雑な多因子相互作用

- **Classification Task**
Activity (Active or Inactive)
- **Regression Task**
LogGI50 value

GI50: 50%の細胞増殖阻害に
必要な化合物濃度

**NCI Human Tumor Cell
Line Growth Inhibition
Assay (PubChem AID 1)**

NCIのヒト腫瘍細胞株の
増殖阻害アッセイ

Active (2,814)			Activity	Score	LogGI50_M
Structure	CID	SID			
	5298	121832	Active	67	-8
	363173	493713	Active	43	-6.5871
	399631	530868	Active	51	-7.0678
Inactive (48,922)			Activity	Score	LogGI50_M
Structure	CID	SID			
	390324	521601	Inactive	0	-4
	390311	521588	Inactive	0	-4
	390312	521589	Inactive	4	-4.214

Use Case ① 定量的活性/物性相關 (QSAR/QSPR)

Standard ML

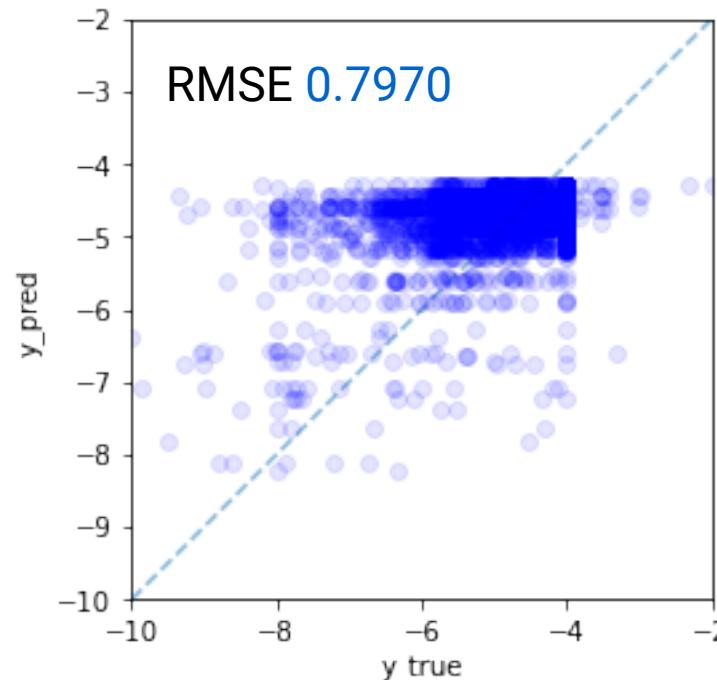
ExtraTrees
w/ ECFP6(1024)

- **Classification Task** Activity (Active or Inactive)

95.079%

- **Regression Task**

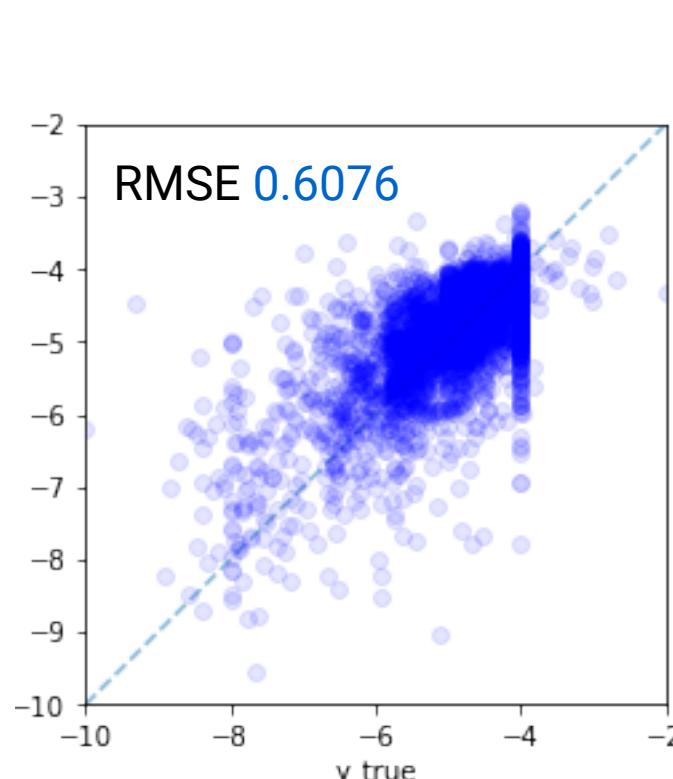
LogGI50 value



GNN

ChemProp
(Directed MPNN)

95.604%



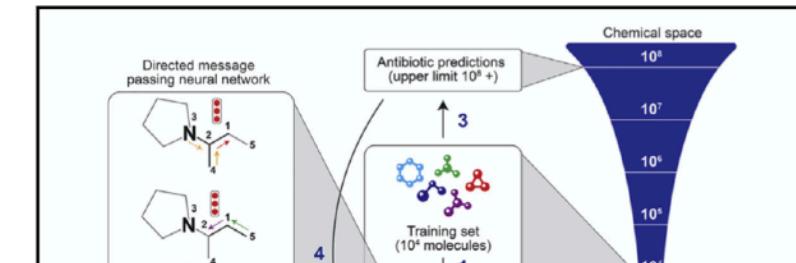
ChemProp (Yang et al, 2019)

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.)

nature

NEWS | 20 February 2020

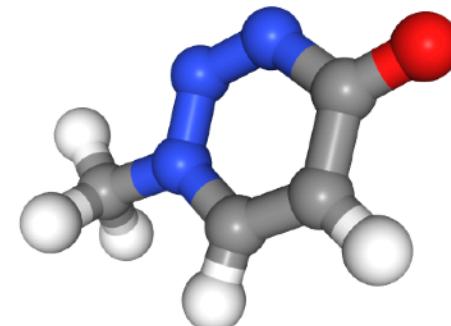
Powerful antibiotics discovered using AI

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

Jo Marchant

Use Case ② 量子化学計算の近似

input



	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

output

- 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
- :

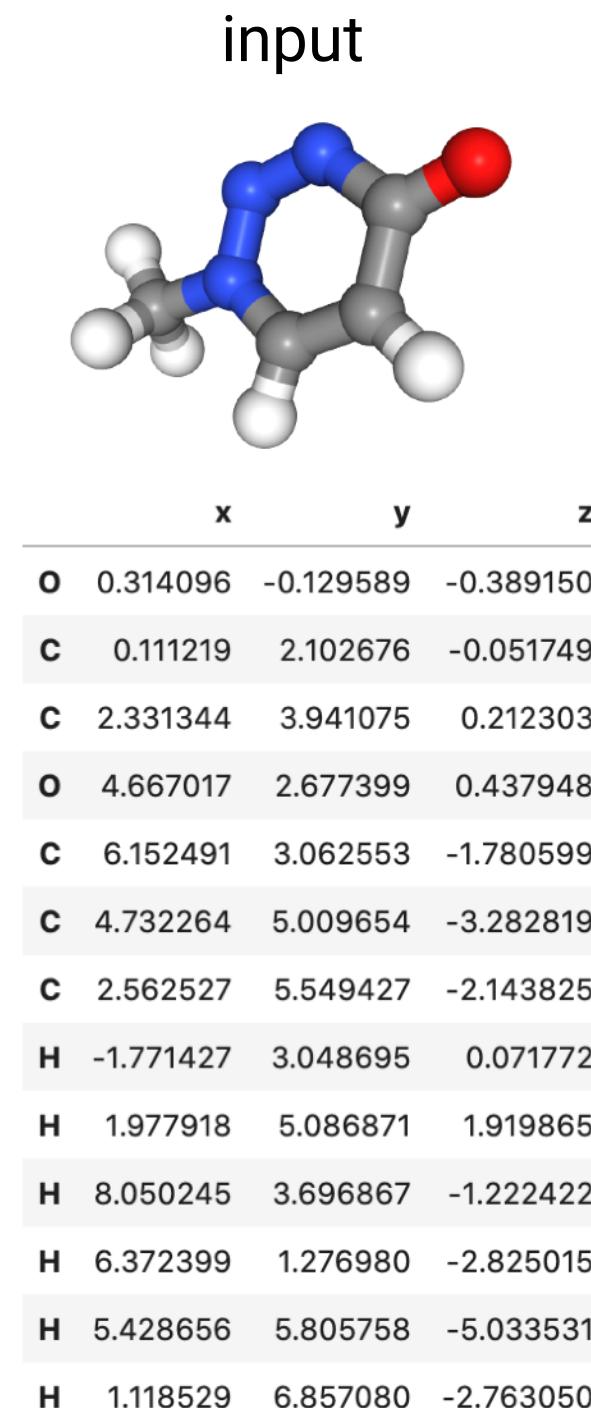
例) 一電子版のSchrödinger方程式
(Kohn-Sham方程式)の求解

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

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

	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 量子化学計算の近似



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

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

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

output

- 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
- ⋮

	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 量子化学計算の近似

 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/

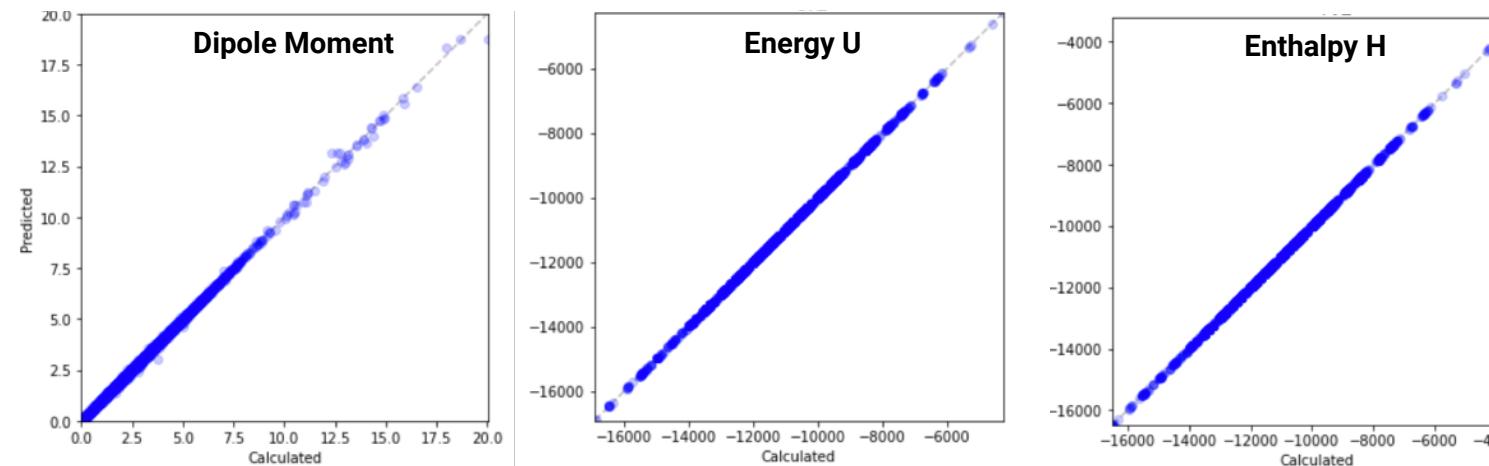
Name	↑↓ Quality	↑↓ Data Points
ANI-1	DFT	22,057,374
ANI-1x	DFT	4,956,005
GDML	DFT, CCSD, CCSD(T)	3,875,468
Solvated Protein Fragments	DFT	2,731,180
ISO-17	DFT	640,982
ANI-1ccx	CCSD(T)*	489,571
SN2 Reactions	DFT	452,709
A Benchmark Data Set for Hydrogen Combustion	wB97X-V/cc-pVTZ	361,803
TensorMol Water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

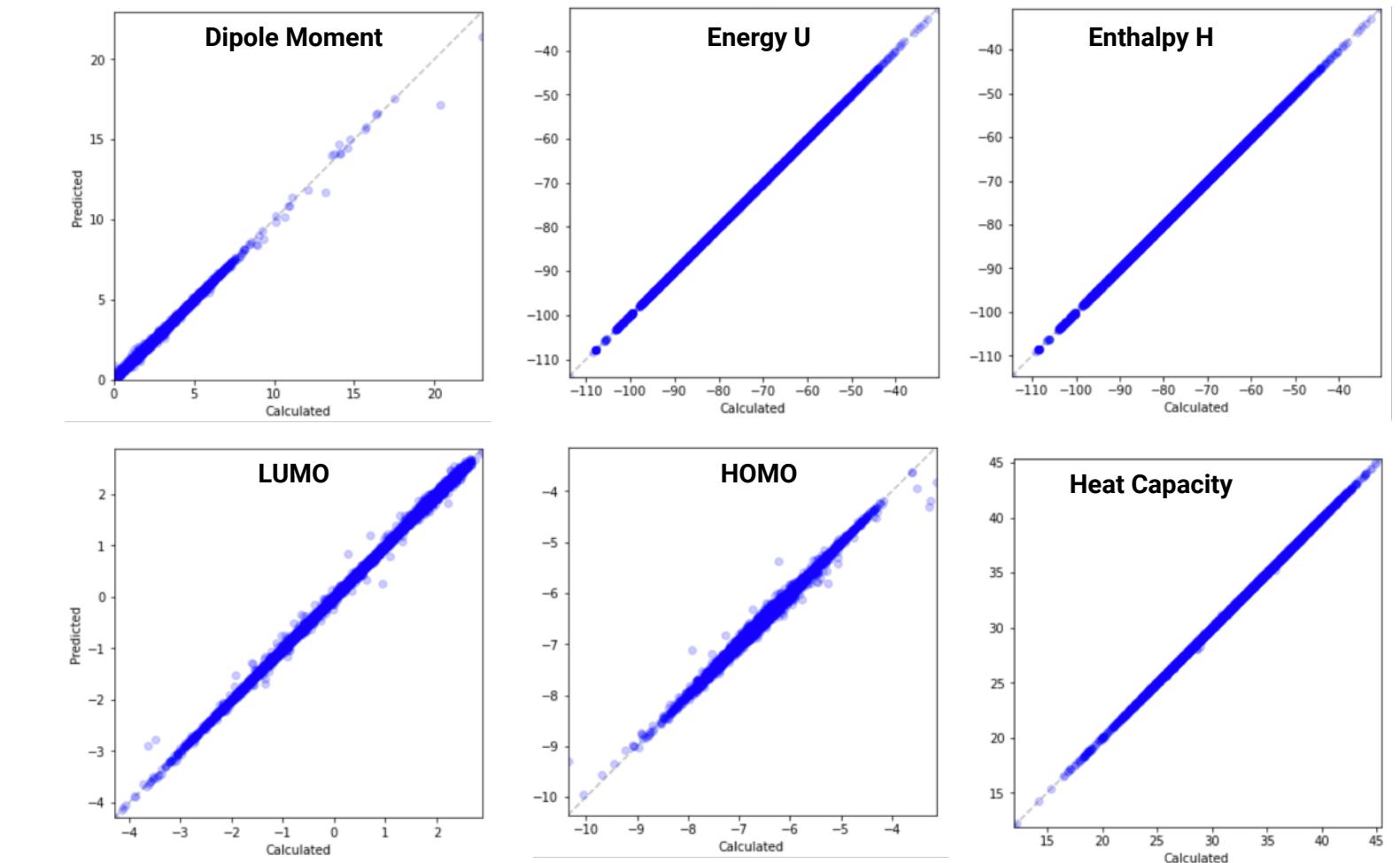
Use Case ② 量子化学計算の近似

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)



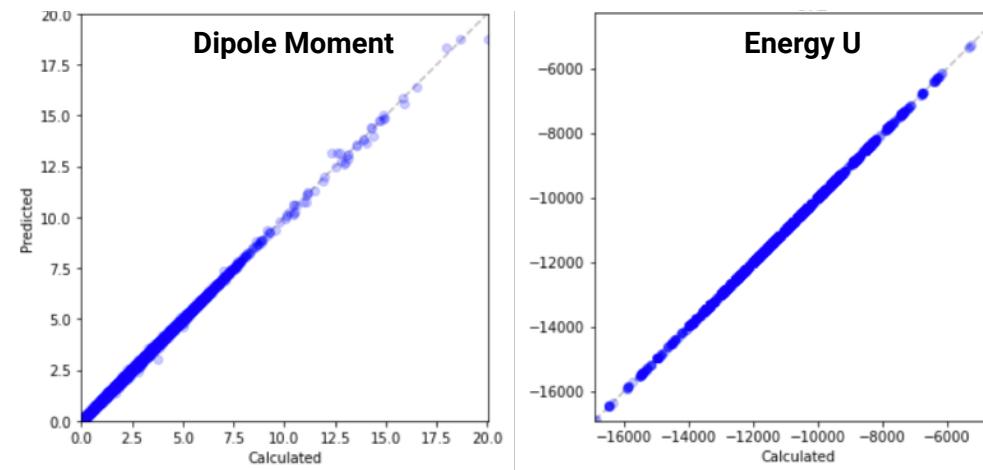
真値(x軸) vs 予測値(y軸) by DimeNet (Klicpera et al, 2020)



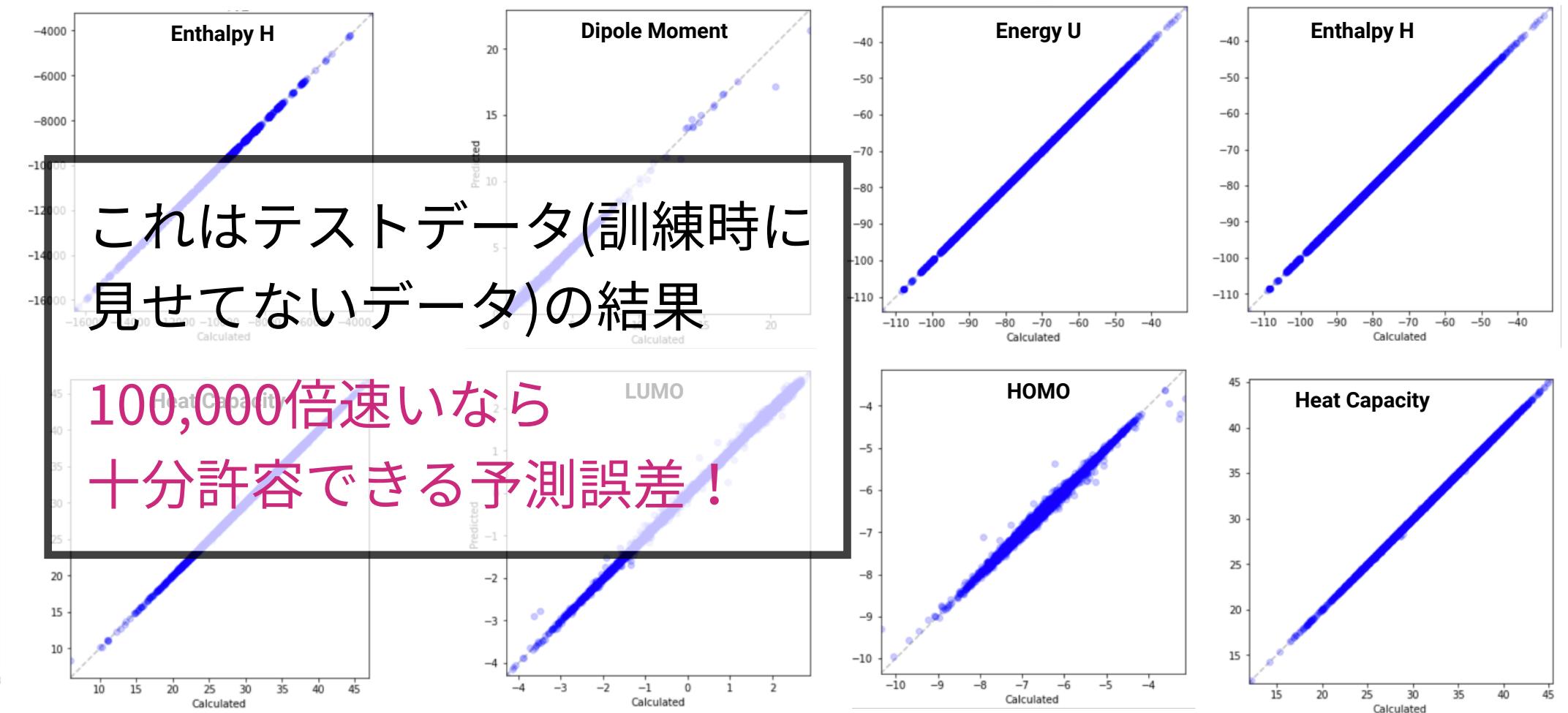
Use Case ② 量子化学計算の近似

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)



Use Case ③ 分子表現の生成

- 表現学習の出力に微分可能なDecoderをつければ分子グラフや分子構造の生成ができる
- 与えられたデータに似た構造を生成するのはそれほど難しくなく生成結果の評価が課題
- 文字列表現(SMILES記法)からの生成に対する優位性が示唆されている

Deep Graph Generators: A Survey

FAEZEH FAEZ¹, YASSAMAN OMMI², MAHDIEH SOLEYMANI BAGHSAAH¹, AND HAMID R.
RABIEE¹, (Senior Member, IEEE)

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

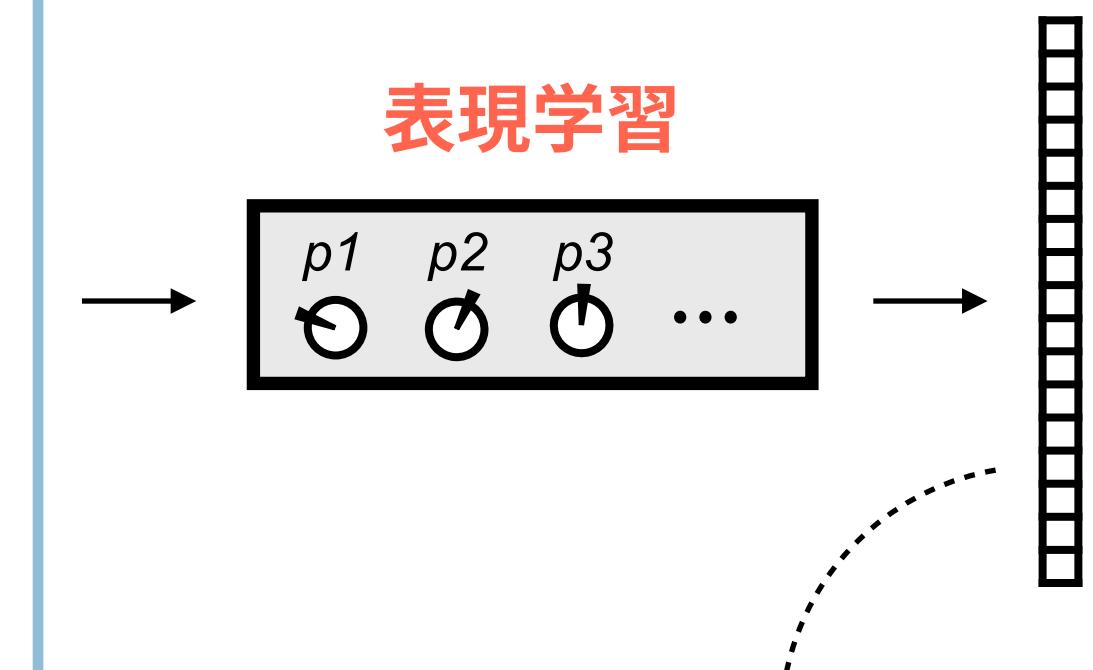
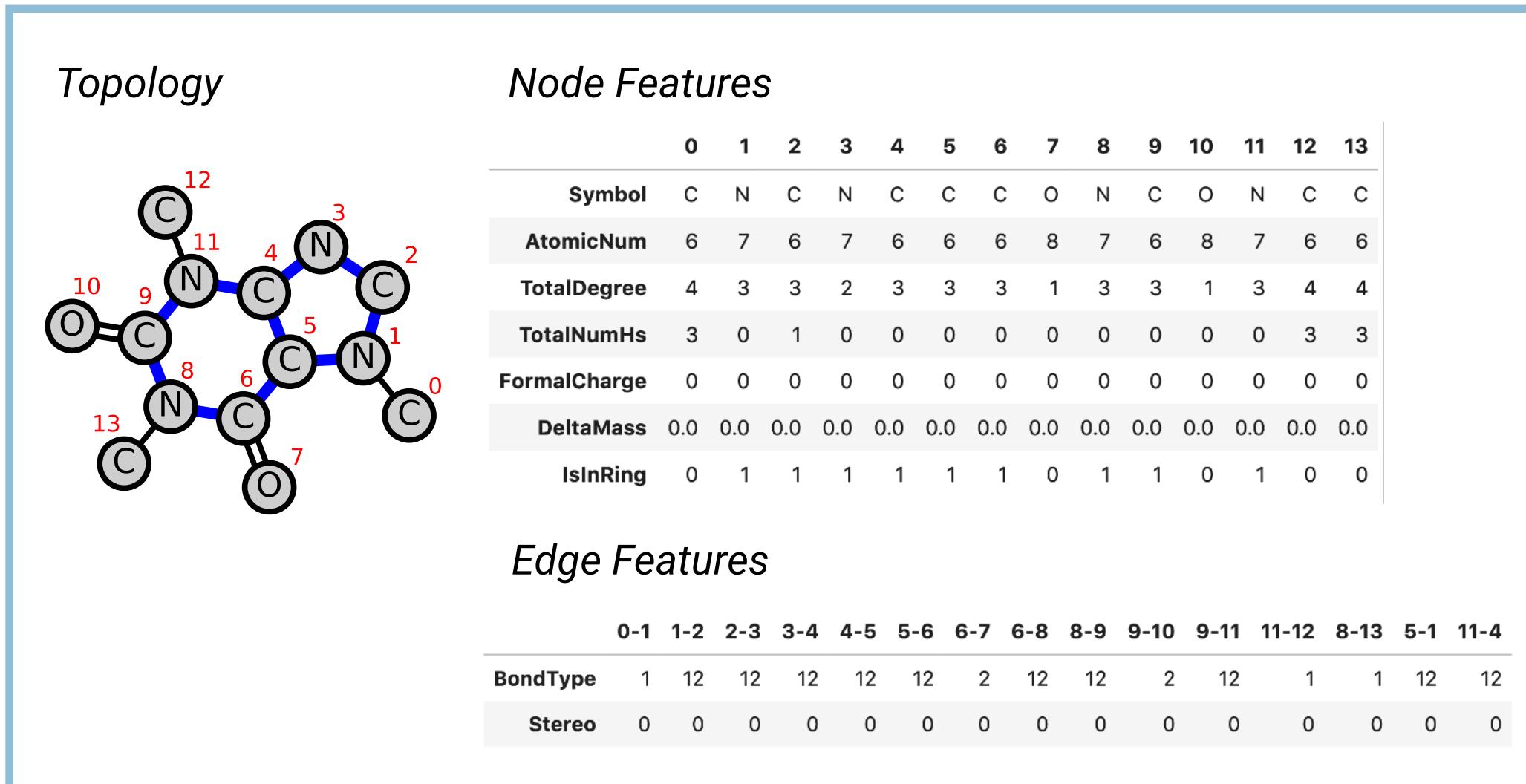
²Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran

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

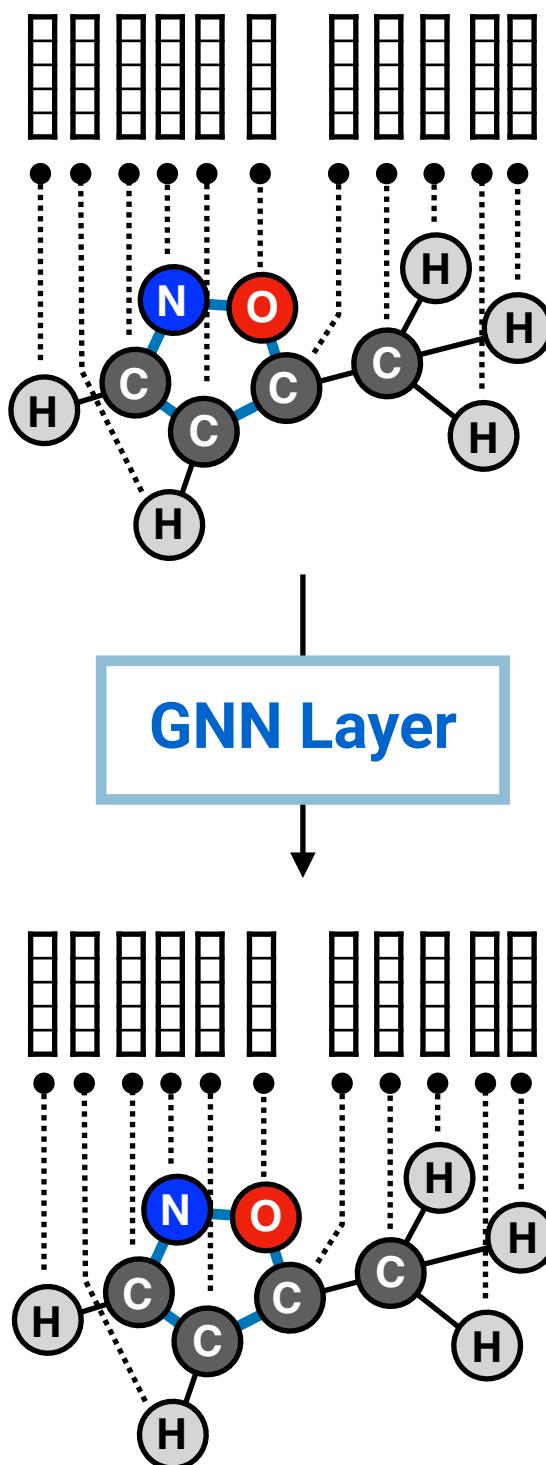
Category	Key Characteristic	Publications
Autoregressive DGGs	Adopting a sequential generation strategy, either node-by-node or edge-by-edge	[1]–[26]
Autoencoder-Based DGGs	Making the generation process dependent on latent space variables	[14]–[19], [27]–[39]
RL-Based DGGs	Utilizing reinforcement learning 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]

グラフの表現学習

分子グラフに付与された情報から汎用の「良い特徴量」を表現学習によって合成したい！



GNNの仕組み：基本処理は頂点特徴量の書き換え



頂点特徴量の書き換えを最も単純にやると例えば…

- ① 書き換えに使う
Neural Netを用意

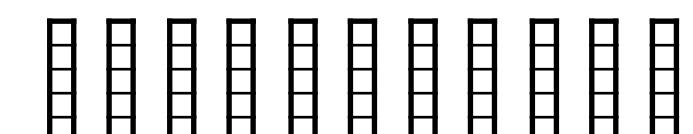
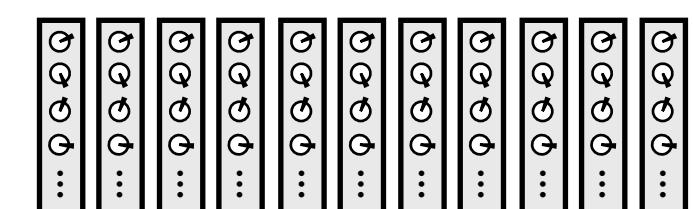
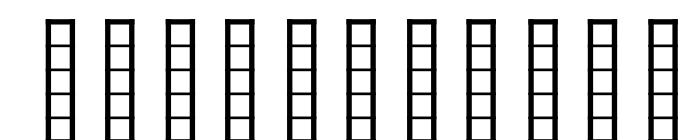
- ② 各々の頂点特徴量に
独立に適用して書き換え

x



いじるパラメタ
はこれだけ

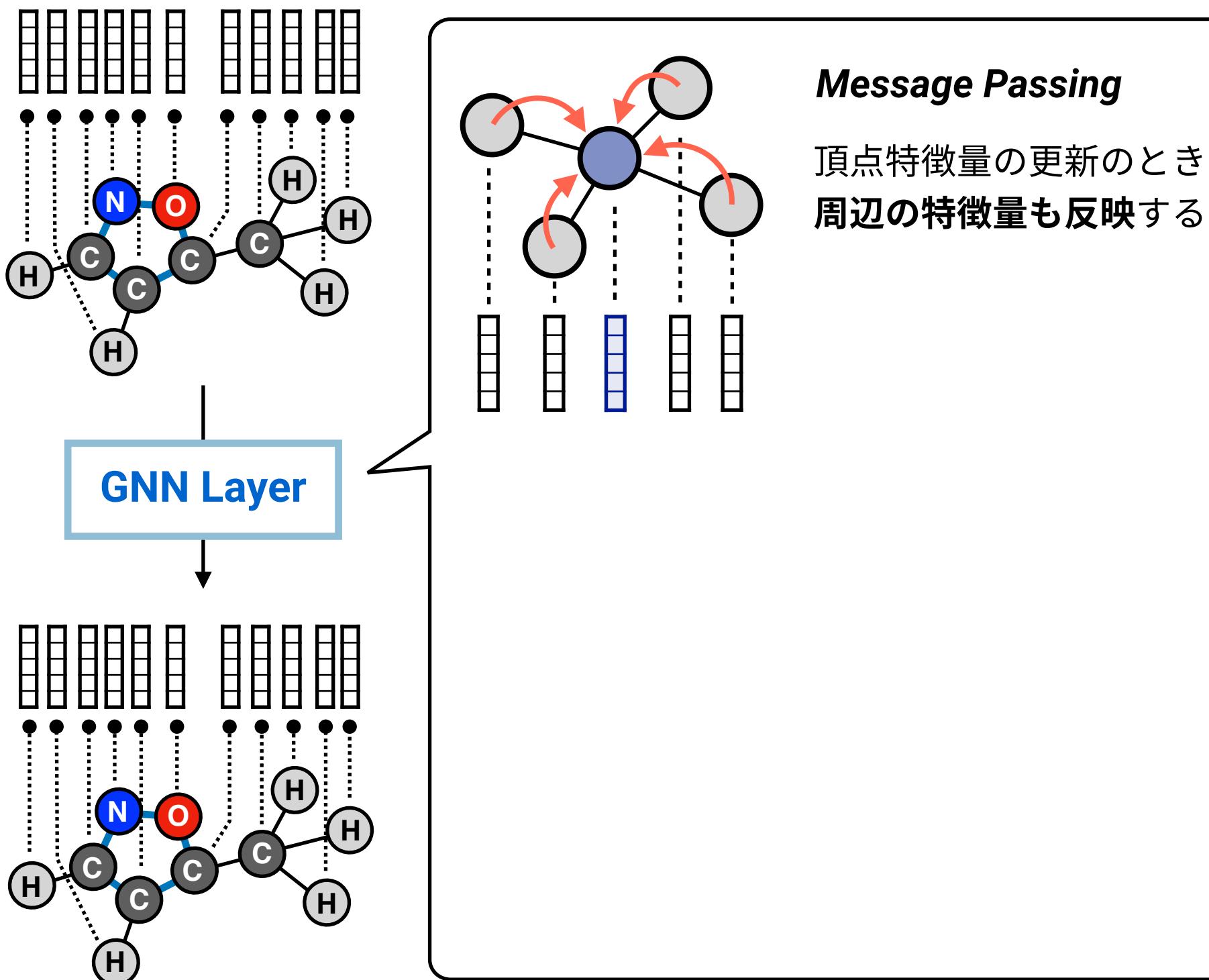
z



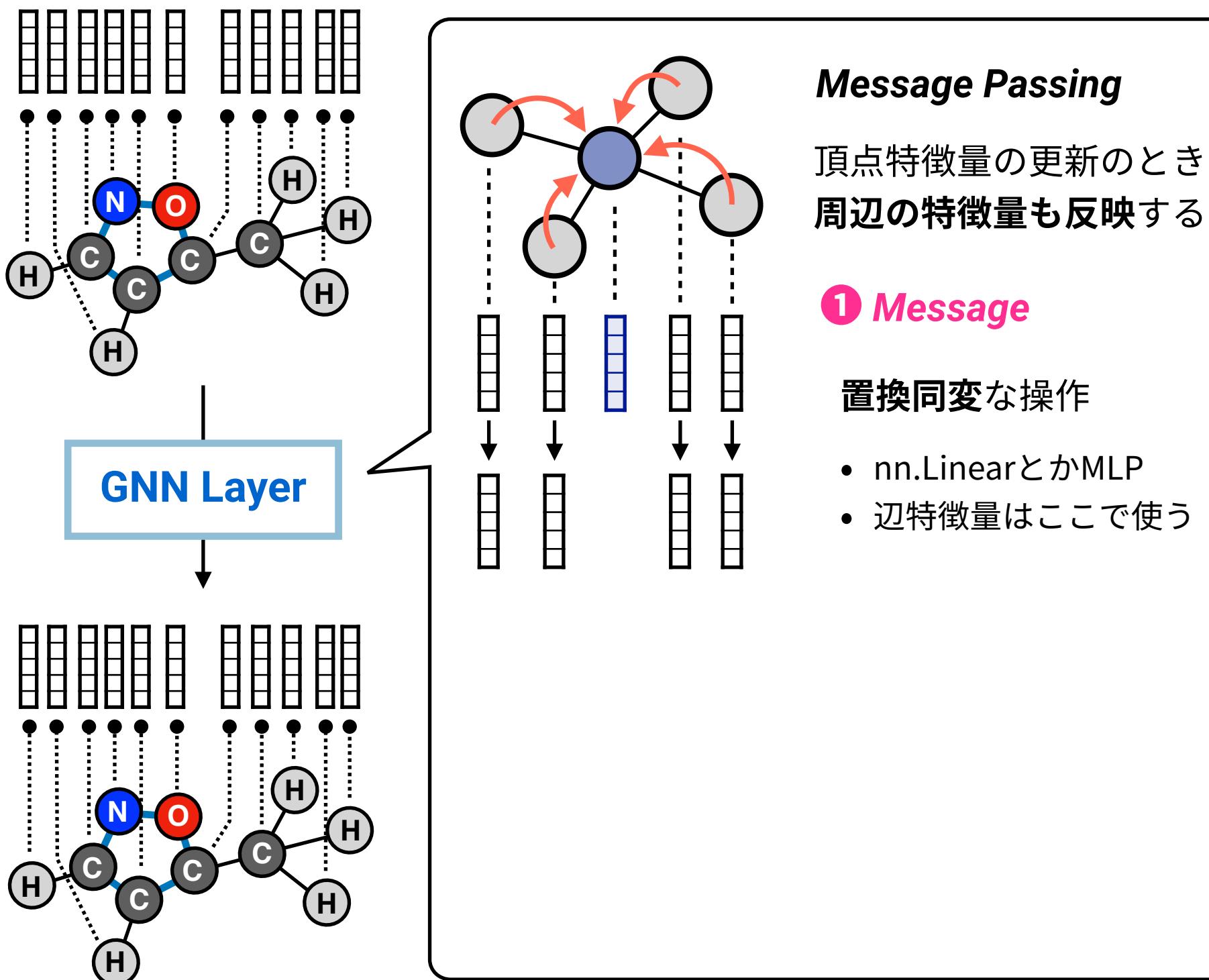
ただし、これだと
原子間の相互作用が一切考慮
されないので書き換えの際
「周辺の特徴量を混ぜる」

→ "Message Passing"

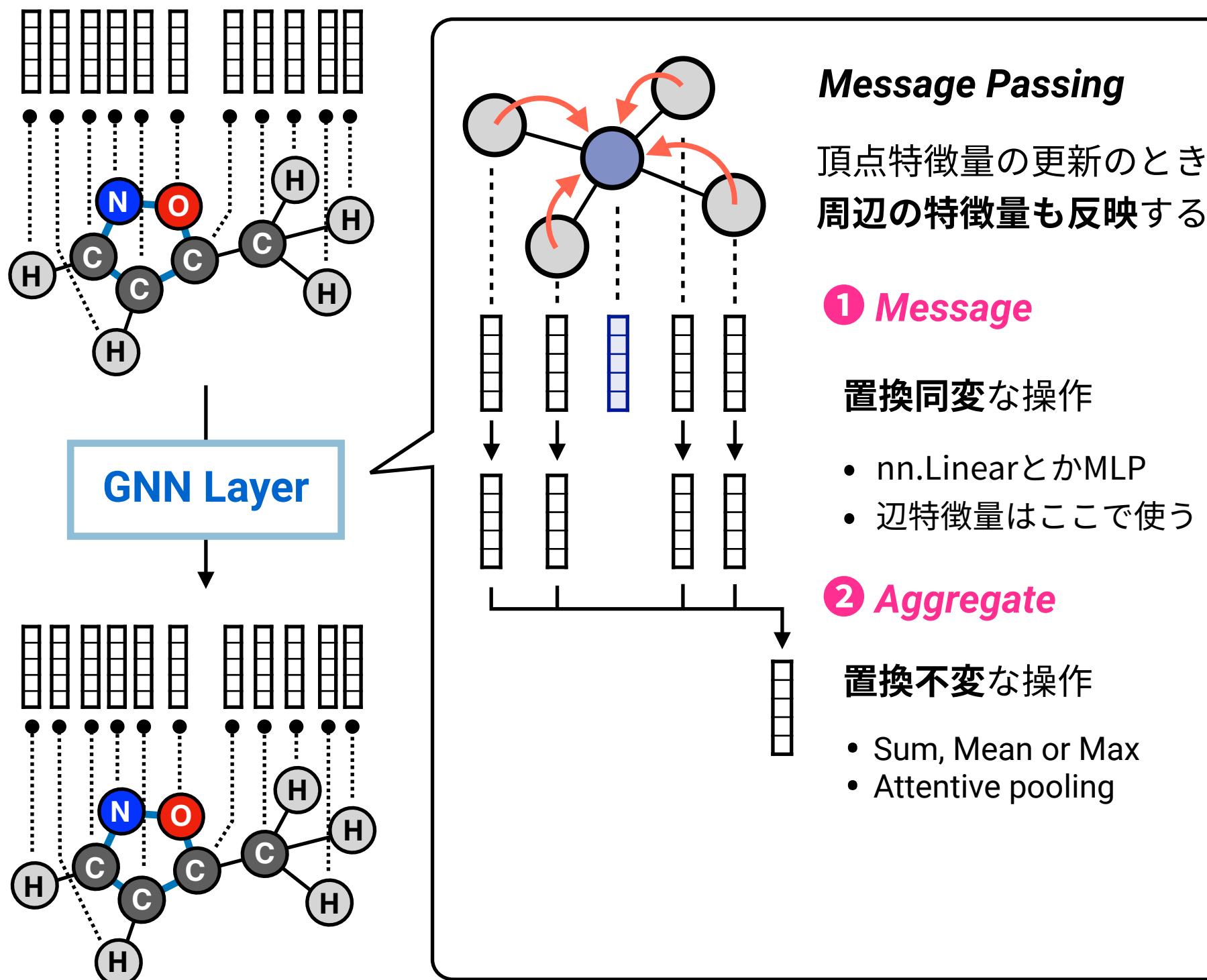
GNNの仕組み : Message Passing



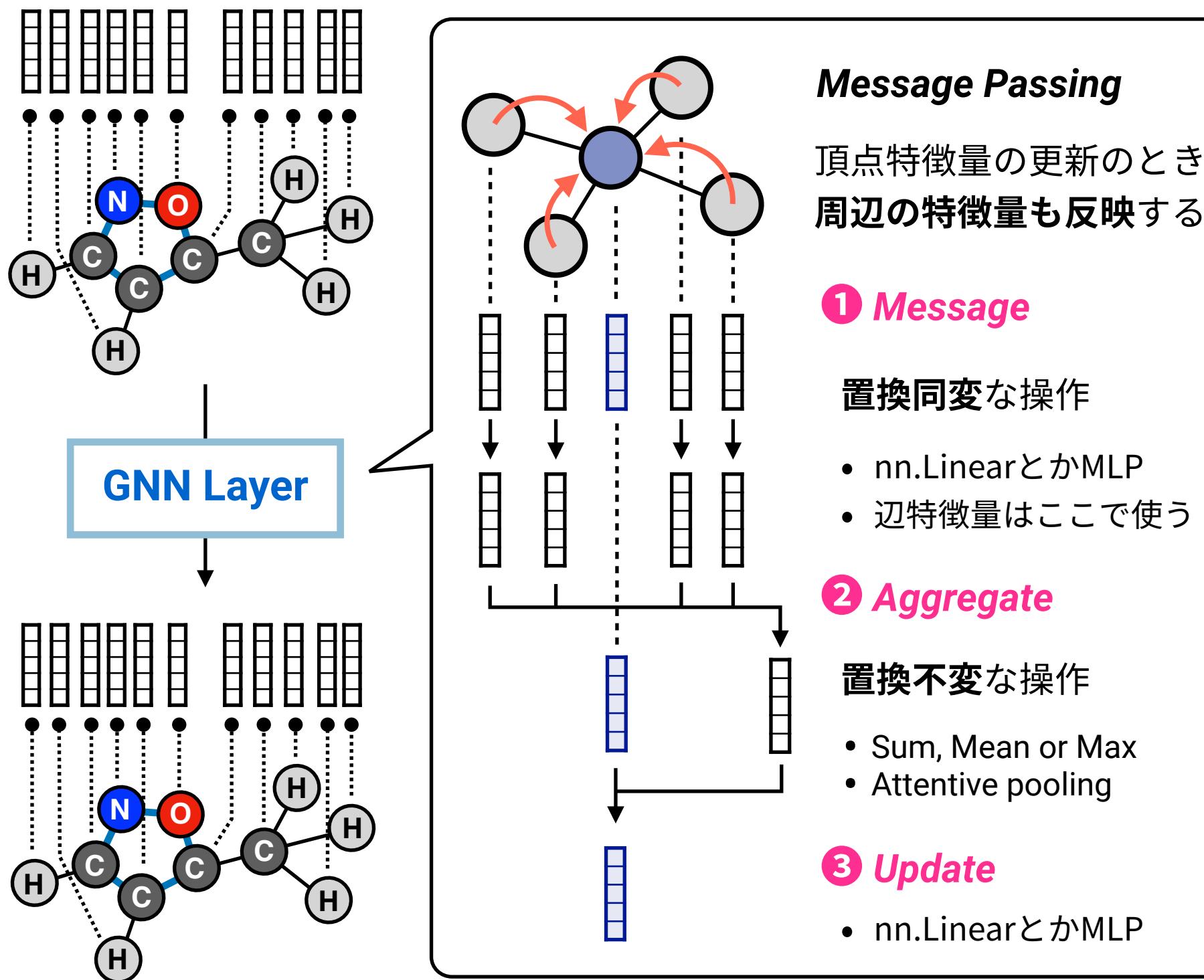
GNNの仕組み : Message Passing



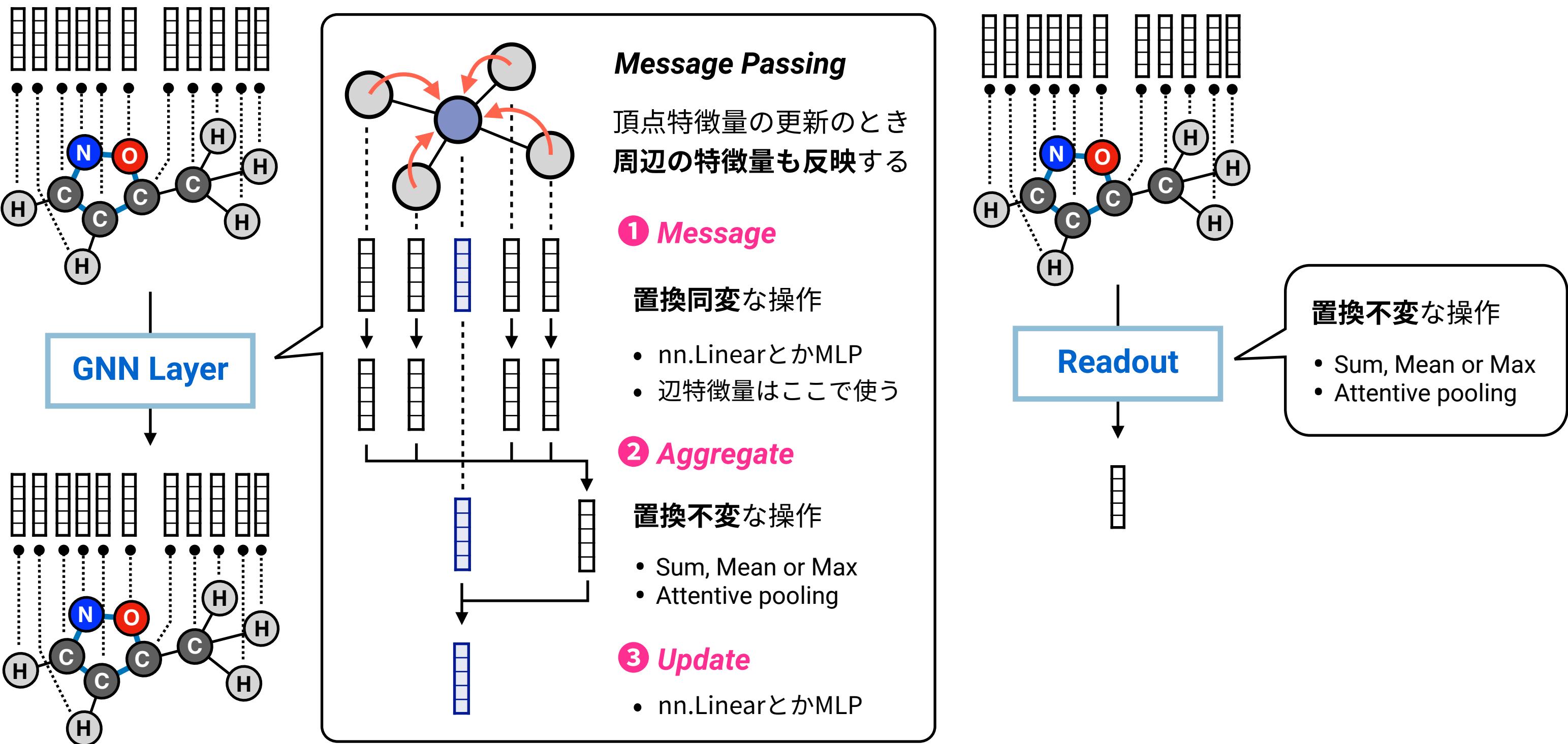
GNNの仕組み : Message Passing



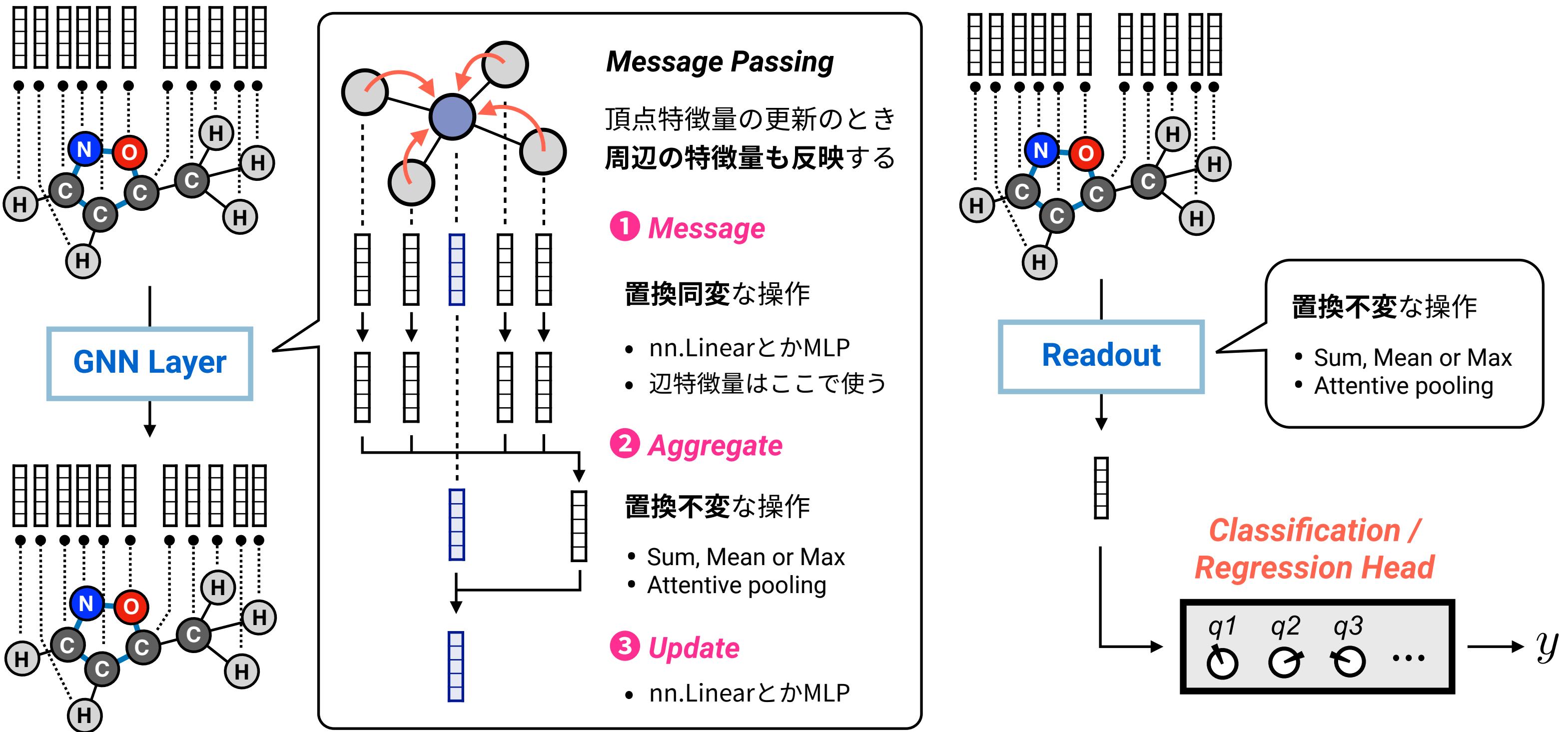
GNNの仕組み : Message Passing



GNNの仕組み : Message Passing

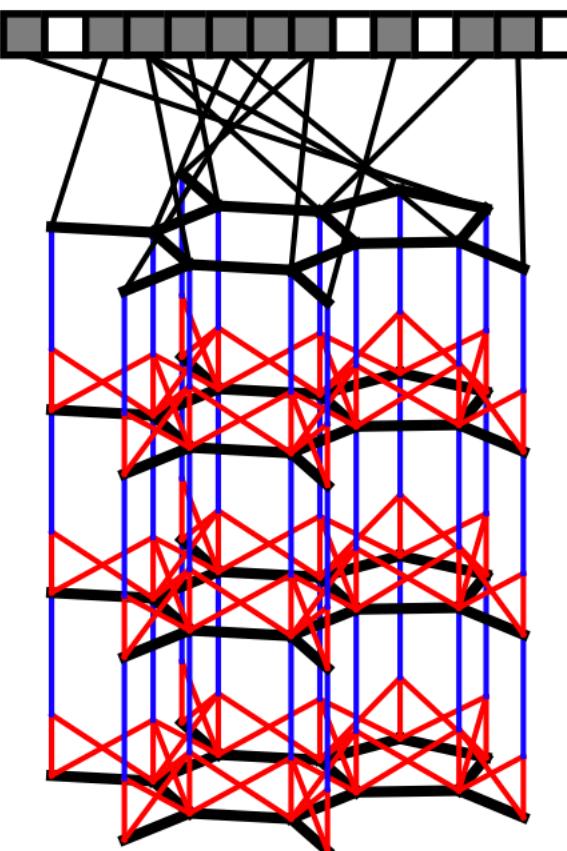


GNNの仕組み : Message Passing



温故知新：ECFPとNeural Graph Fingerprint(初期GNN)

- Neural Graph Fingerprint: 最初期に提案されたGNNの一つ
- ケモインフォマティクス分野で分子グラフから得られる記述子ECFP(Circular Fingerprint)の計算をパラメタを持つ微分可能な演算で書き直したもの



Algorithm 1 Circular fingerprints

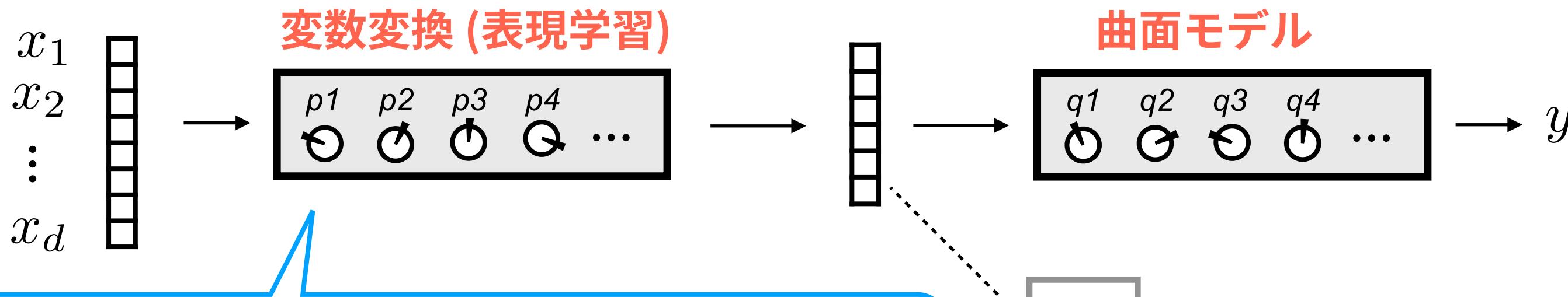
```
1: Input: molecule, radius  $R$ , fingerprint  
length  $S$   
2: Initialize: fingerprint vector  $\mathbf{f} \leftarrow \mathbf{0}_S$   
3: for each atom  $a$  in molecule  
4:    $\mathbf{r}_a \leftarrow g(a)$             $\triangleright$  lookup atom features  
5:   for  $L = 1$  to  $R$             $\triangleright$  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:       $f_i \leftarrow 1$                    $\triangleright$  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.

小サンプル問題の克服：大規模事前学習とその転移



収集可能な何らかの「大規模なデータ」に対して
本来のタスクとは別の「自己教師あり(SSL)」タスク
(Pretextタスク)を設計し変数変換部を事前に獲得する

SSL Pretextタスクの例 (人手の正解ラベルづけが不要)

- 文の単語をランダムに隠してそれを当てる
- 絵を $90^\circ/180^\circ/270^\circ$ にランダム回転させ角度を当てる
- 分子構造の一部を隠してそれを当てる

理想 Few-shot / Zero-shotでの転移

「良い表現」が得られれば曲面モデルは
シンプル(線形)で良く 小サンプルでOK

現実 大規模データの取得が困難…

画像・音声・テキストなど華々しい成功例
のケースと違い、自然現象を扱う分野では
大規模データの取得が現状難しい…

事前学習は深層学習を実用ツールに変えた (ImageNet, BERT,...)

"Foundation Model"と呼んじゃおう！とStanfordが研究センターまで設立し議論を呼んだ…

arXiv.org > cs > arXiv:2108.07258

Search...
Help | Advanced

Computer Science > Machine Learning

[Submitted on 16 Aug 2021 (v1), last revised 18 Aug 2021 (this version, v2)]

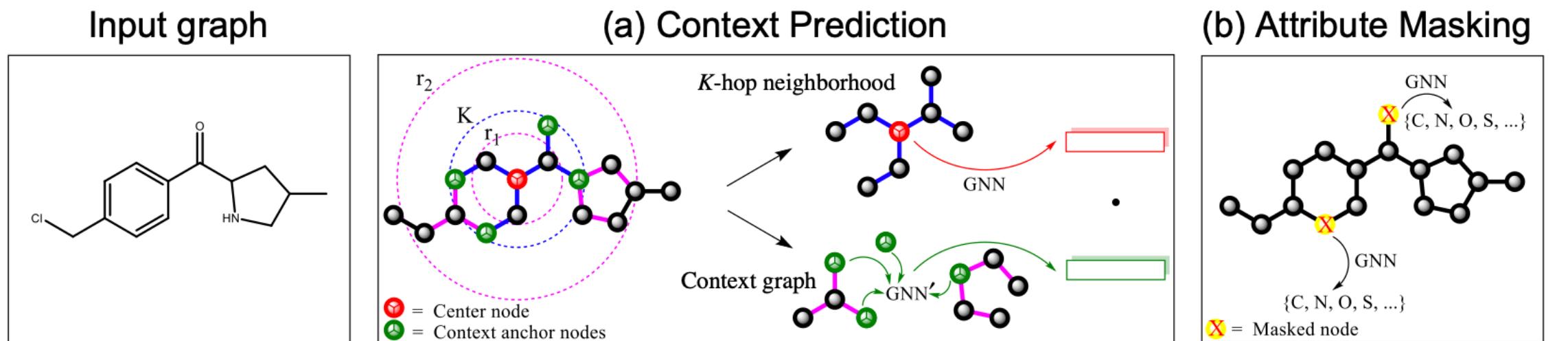
On the Opportunities and Risks of Foundation Models

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang et al. (14 additional authors not shown)

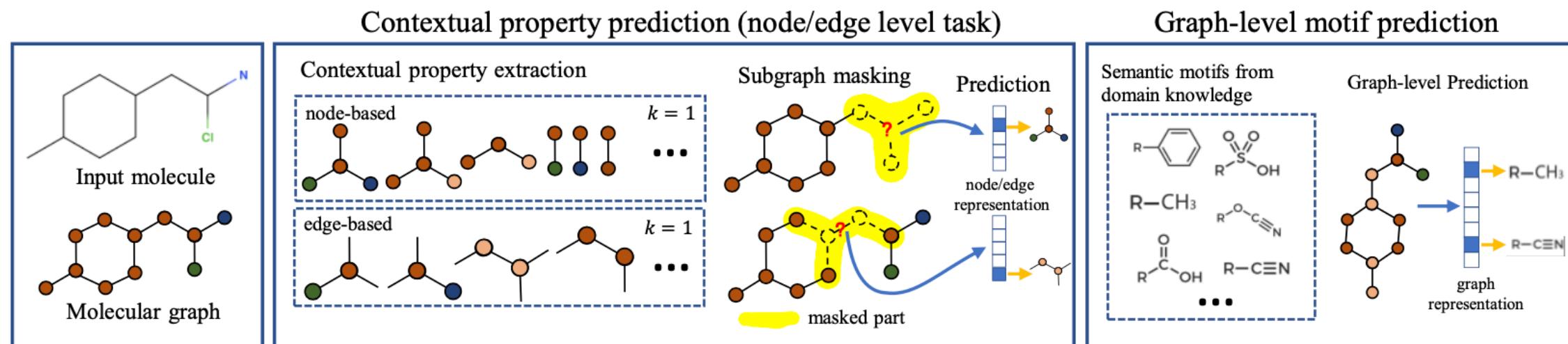
AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character. This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotics, reasoning, human interaction) and technical principles (e.g., model architectures, training procedures, data, systems, security, evaluation, theory) to their applications (e.g., law, healthcare, education) and societal impact (e.g., inequity, misuse, economic and environmental impact, legal and ethical considerations). Though foundation models are based on standard deep learning and transfer learning, their scale results in new emergent capabilities, and their effectiveness across so many tasks incentivizes homogenization. Homogenization provides powerful leverage but demands caution, as the defects of the foundation model are inherited by all the adapted models downstream. Despite the impending widespread deployment of foundation models, we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties. To tackle these questions, we believe much of the critical research on foundation models will require deep interdisciplinary collaboration commensurate with their fundamentally sociotechnical nature.

分子表現の事前学習と転移学習

- GNNによる分子表現の大規模事前学習(特にSSLタスク設計)は現在最もアツイ話題の一つ
- SSLは教師ラベルづけ不要 + 分子は多分野に分類・回帰・生成など幅広い下流タスクを持つ



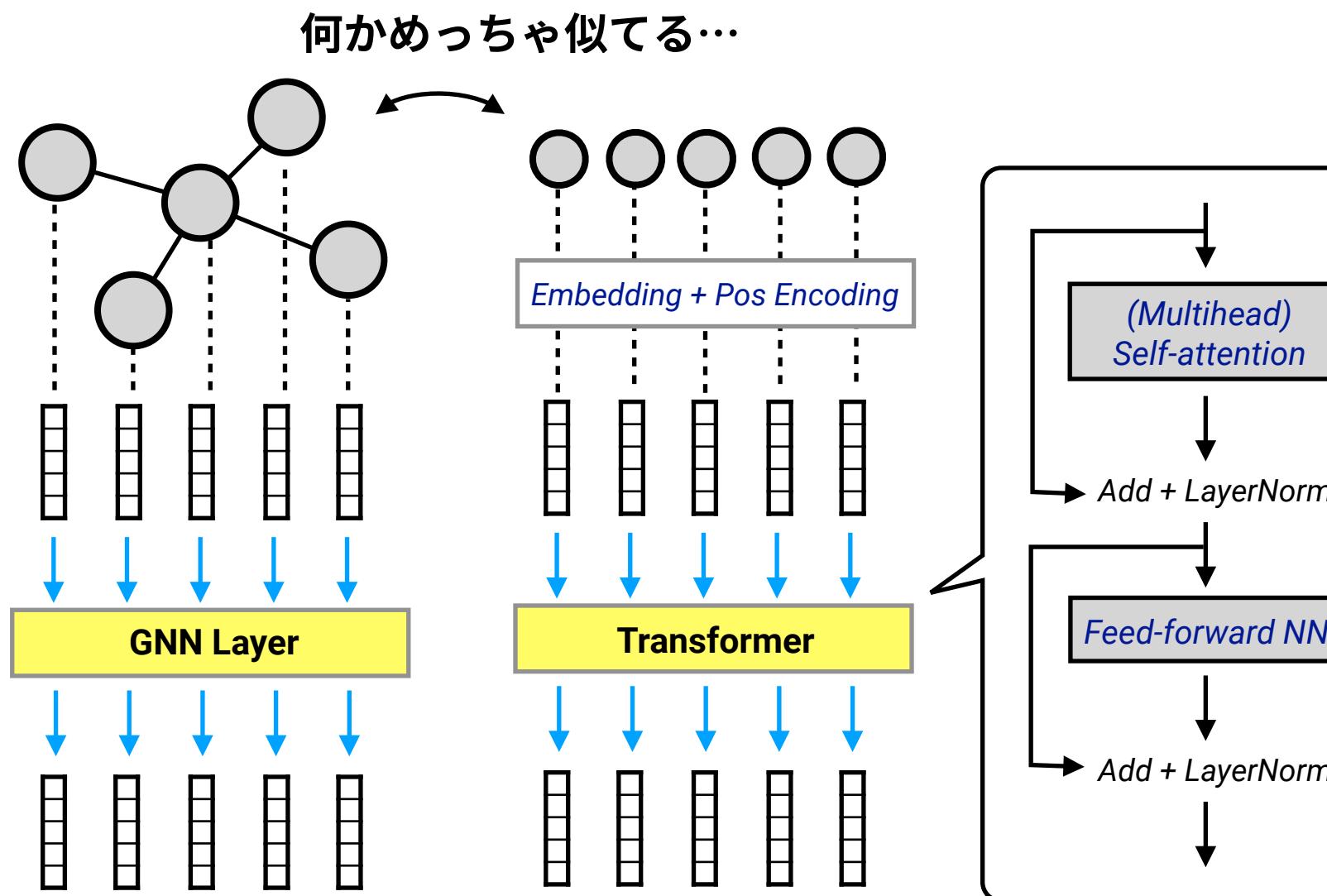
*Strategies for Pre-training
Graph Neural Networks*
Hu, Liu, Gomes, Zitnik, Liang,
Pande, Leskovec (ICLR 2020)
<https://arxiv.org/abs/1905.12265>



*Self-Supervised Graph
Transformer on Large-Scale
Molecular Data*
Rong, Bian, Xu, Xie, Wei, Huang,
Huang (NeurIPS 2020)
<https://arxiv.org/abs/2007.02835>

GATとTransformer型GNN

- Graph Attention Network (GAT) = 頂点特徴量の更新にAttentionを入れた基本的GNN
- Transformerはトポロジ制約のない**GAT**の変種とみなせる
- 逆にTransformer型のSelf-AttentionをGNNにもちこむこともできる



関心

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は分子タスクも変えるか？

A Generalization of Transformer Networks to Graphs

Dwivedi & Bresson (2020) <https://arxiv.org/abs/2012.09699>

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

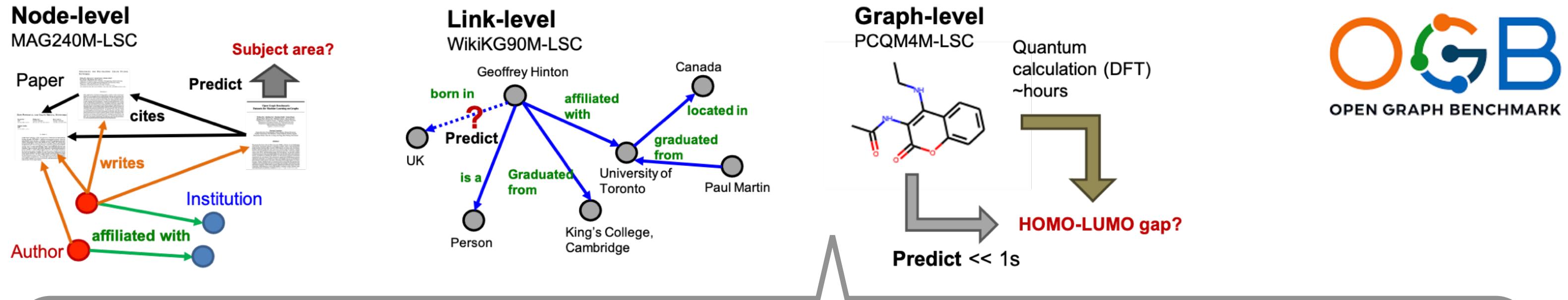
Zhang et al (2020) <https://arxiv.org/abs/2001.05140>

Do Transformers Really Perform Bad for Graph Representation?

Ying et al (2021) <https://arxiv.org/abs/2106.05234>

→ KDDCup 2021のGraph-levelタスクで優勝

OGB Large-Scale Challenge (KDDCup 2021)

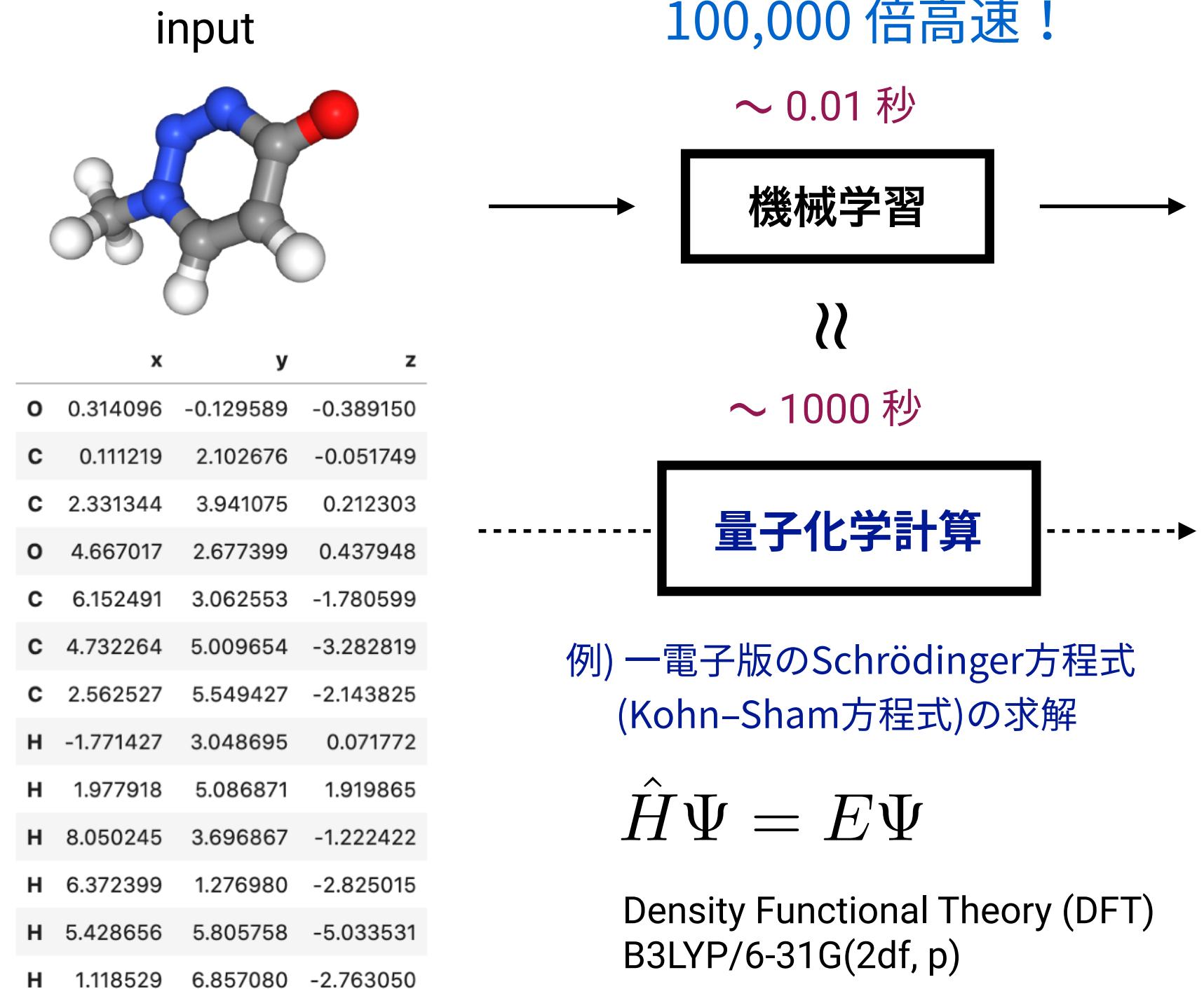


2Dの分子グラフから量子化学計算で求めたHOMO-LUMOギャップを予測するタスク

データセット：PubChemQCから3,803,453グラフ (cf. QM9は133,885グラフ)

順位	Test MAE	手法
第1位	0.1200 (eV)	10 × GNNs (12層 Graphomer) + 8 × ExpC*s (5層 ExpandingConv)
第2位	0.1204 (eV)	73 × GNNs (11層 LiteGEMConv + SSL事前学習)
第3位	0.1205 (eV)	20 × GNNs (32層 GNN + Noisy Nodes)

GNNの汎用性：3Dの幾何構造も同じやり方で扱える！



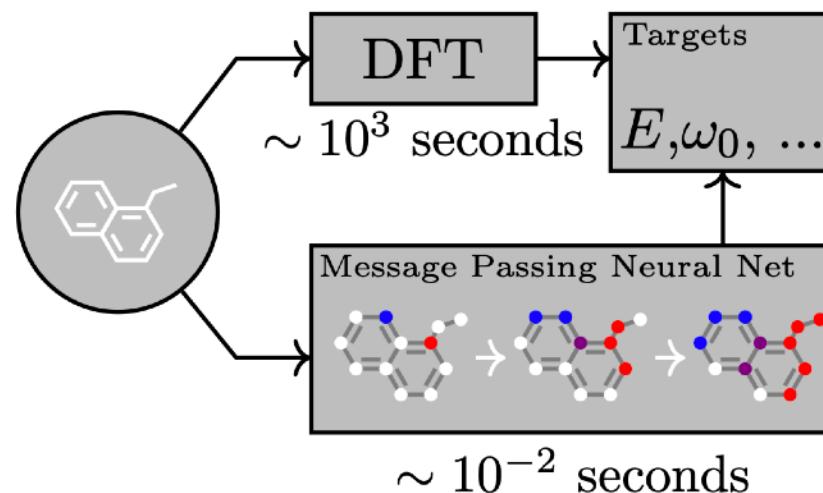
output

- 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
- ⋮

	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

GNN × 3D点幾何(量子化学計算)

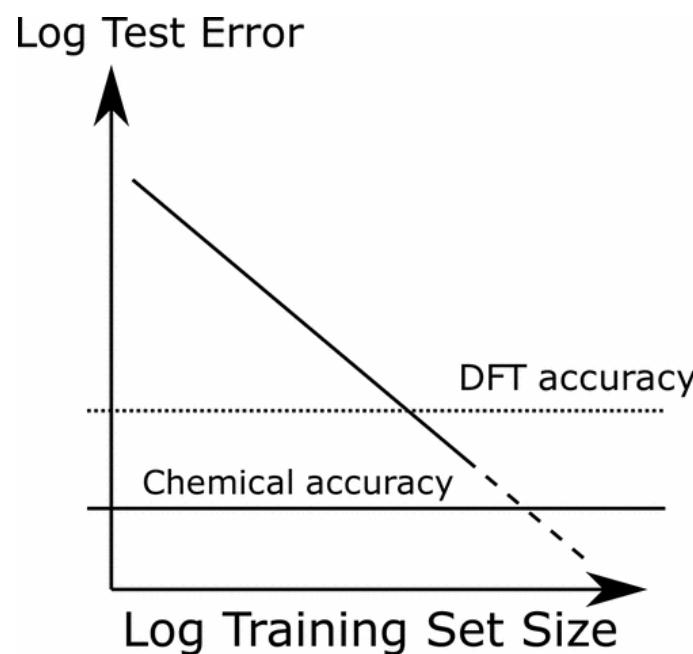
Message Passingとして統一的に既存のGNNを説明したGoogle論文と続報で対象タスクに



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



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

JCTC

Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

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}

この時点では結合長(原子間距離)を辺特徴量に入れただけ

ケモインフォマティクスの頂点・辺特徴

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?

ベクトル量だが離散ラベルとして
部分グラフ同型判定に使う



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

連続量ラベルを
辺特徴量に使う

幾何的深層學習

GNNは幅広い幾何構造を統一的に扱える枠組み (機械學習のエルランゲン・プログラム!?)

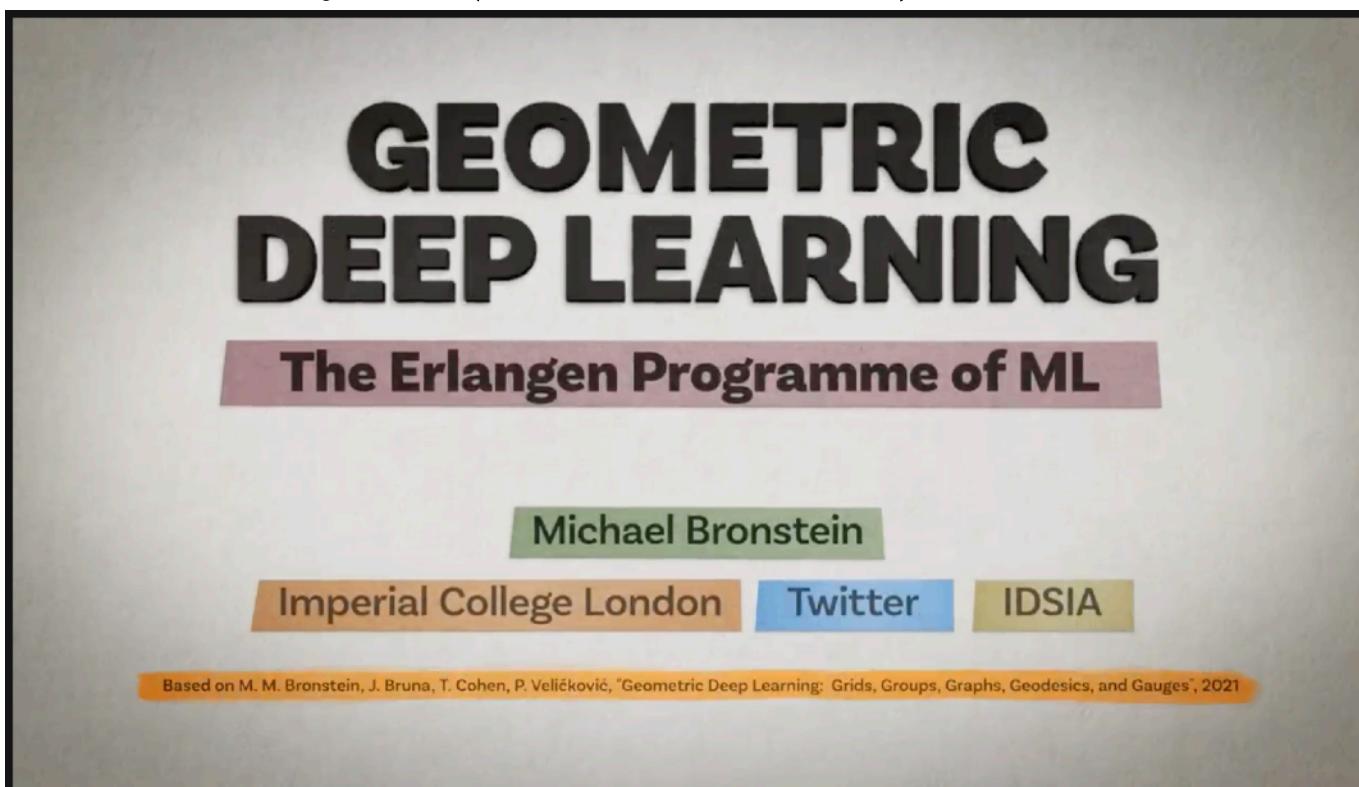
[Submitted on 27 Apr 2021 (v1), last revised 2 May 2021 (this version, v2)]

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

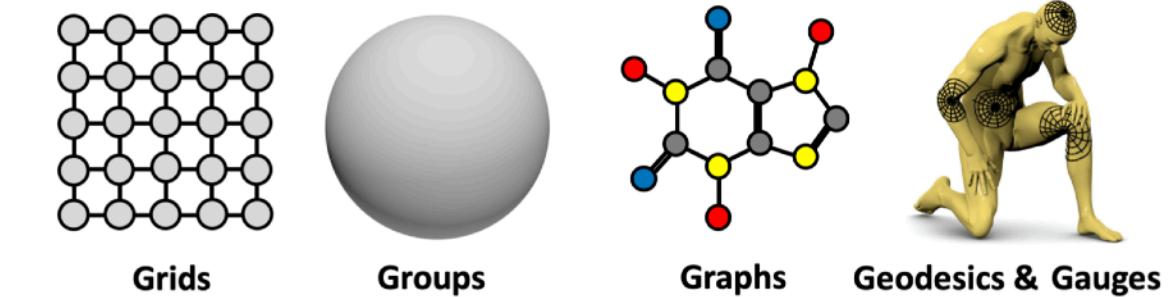
Michael M. Bronstein, Joan Bruna, Taco Cohen, Petar Veličković

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

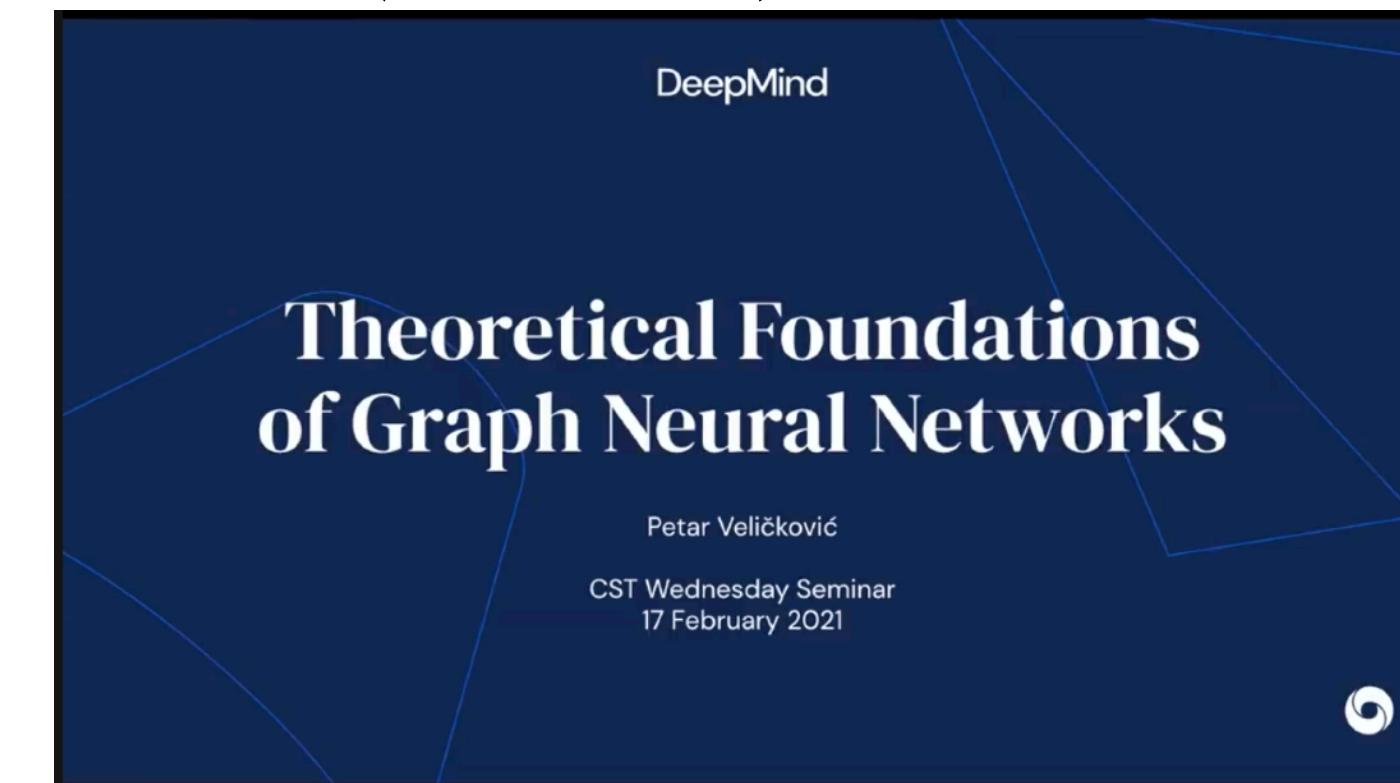
ICLR 2021 Keynote (Michael Bronstein)



<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・同変性の考慮

幾何的GNNの基本要件：原子のxyz座標値をそのまま頂点特徴量にするのは×

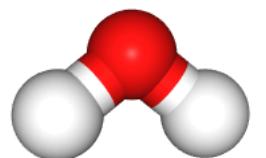
例えば平行移動や回転でxyzは変わるがその分子のエネルギーは変わらない

頂点や辺の特徴量やGNNアーキテクチャのデザインで実現する

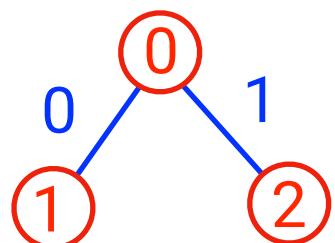
	• ユークリッド群 $E(3)$: 3Dの並進・回転対称性	不变 (invariant)
	• 特殊ユークリッド群 $SE(3)$: $E(3) + 鏡像対称性$	$f(g \cdot x) = f(x)$
$E(3)$ 不变	Schütt et al, SchNet . (2017) https://arxiv.org/abs/1706.08566	
	Unke et al, PhysNet . (2019) https://arxiv.org/abs/1902.08408	同変 (equivariant)
	Klicpera et al, DimeNet++ . (2020) https://arxiv.org/abs/2011.14115	$f(g \cdot x) = g \cdot f(x)$
$SE(3)$ 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	

化学的に既に分かっていることを不要に学習させないために

Water molecule H₂O



A molecular graph (RDKit)



	0	1	2
AtomicNum	8	1	1
TotalDegree	2	1	1
TotalNumHs	0	0	0
FormalCharge	0	0	0
deltaMass	0	0	0
IsInRing	0	0	0

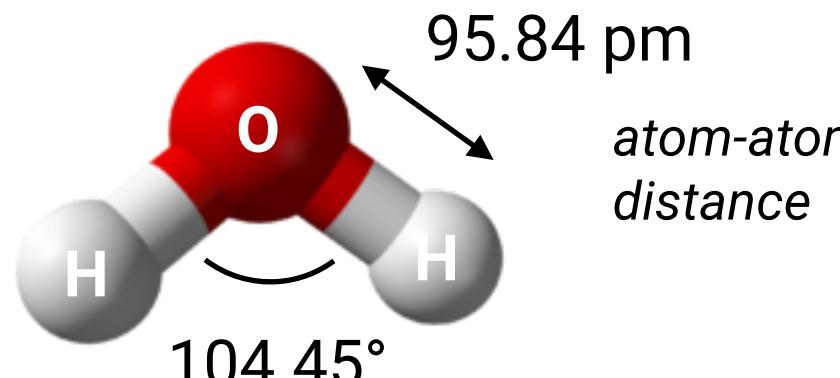
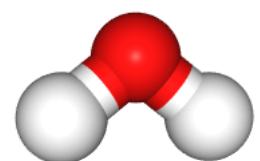
*Atom
Invariants*

	0	1
BondType	0	0
Stereo	0	0

*Bond
Invariants*

化学的に既に分かっていることを不要に学習させないために

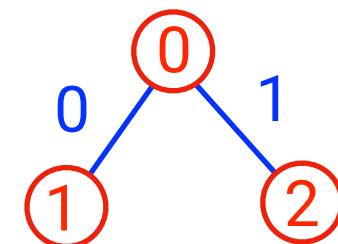
Water molecule H₂O



$$8 + 1 + 1 = 10 \text{ electrons}$$

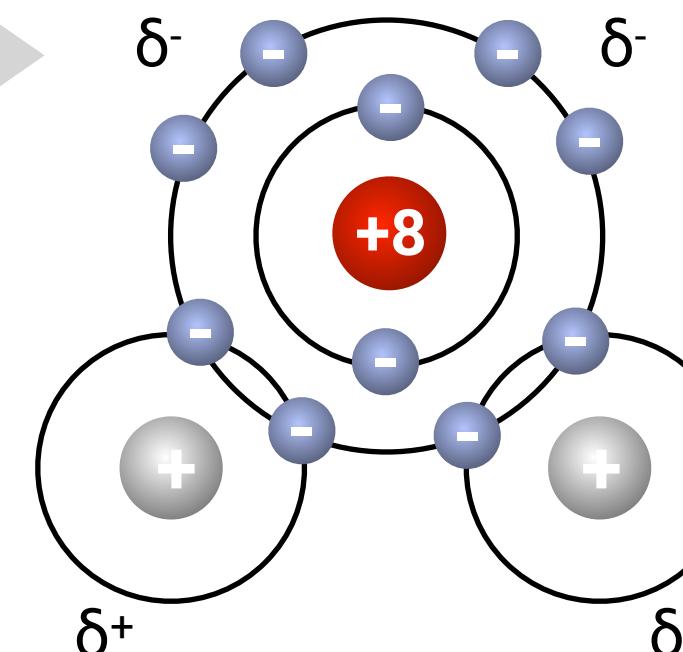
O $1s^2 2s^2 2p^4$
H $1s^1$ H $1s^1$

A molecular graph (RDKit)



AtomicNum	8
TotalDegree	2
TotalNumHs	0
FormalCharge	0
deltaMass	0
IsInRing	0

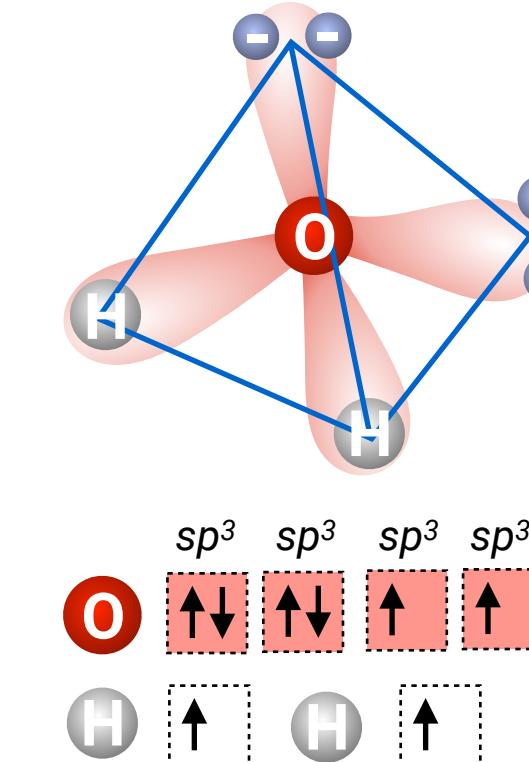
Atom
Invariants



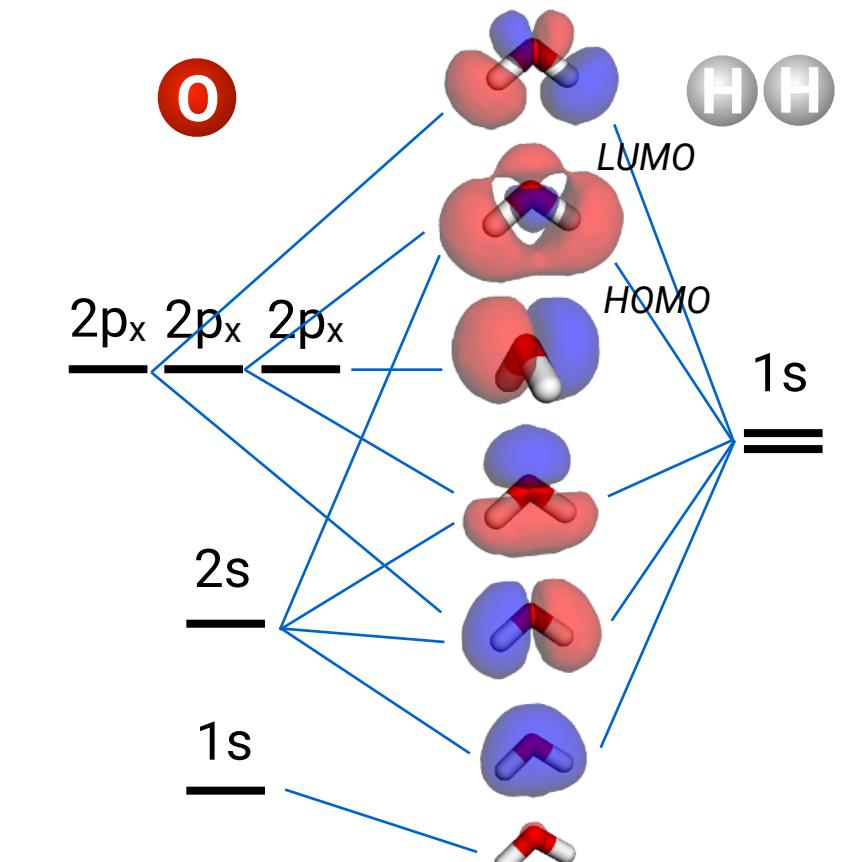
bond angle

BondType	0
Stereo	0

Bond
Invariants

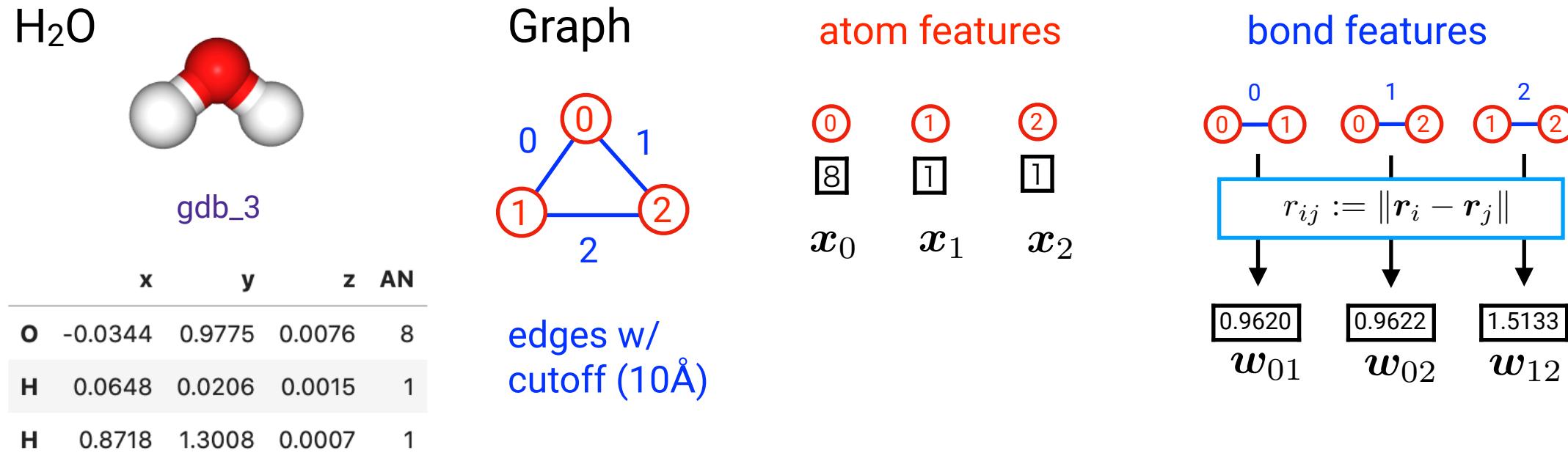


sp³ hybridization



molecular orbitals

SchNet (Schütt et al, 2017): 幾何的GNNの先駆的Standard



SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions.

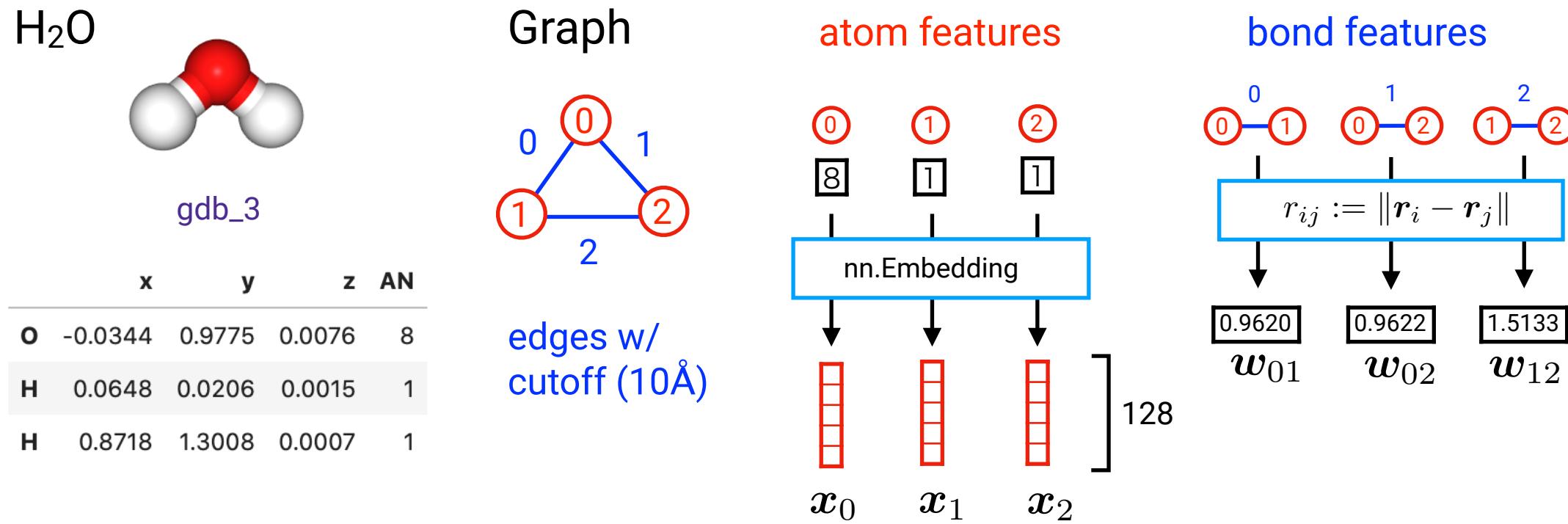
Schütt, Kindermans, Sauceda, Chmiela,

Tkatchenko, Müller

(NIPS 2017)

<https://arxiv.org/abs/1706.08566>

SchNet (Schütt et al, 2017): 幾何的GNNの先駆的Standard



SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions.

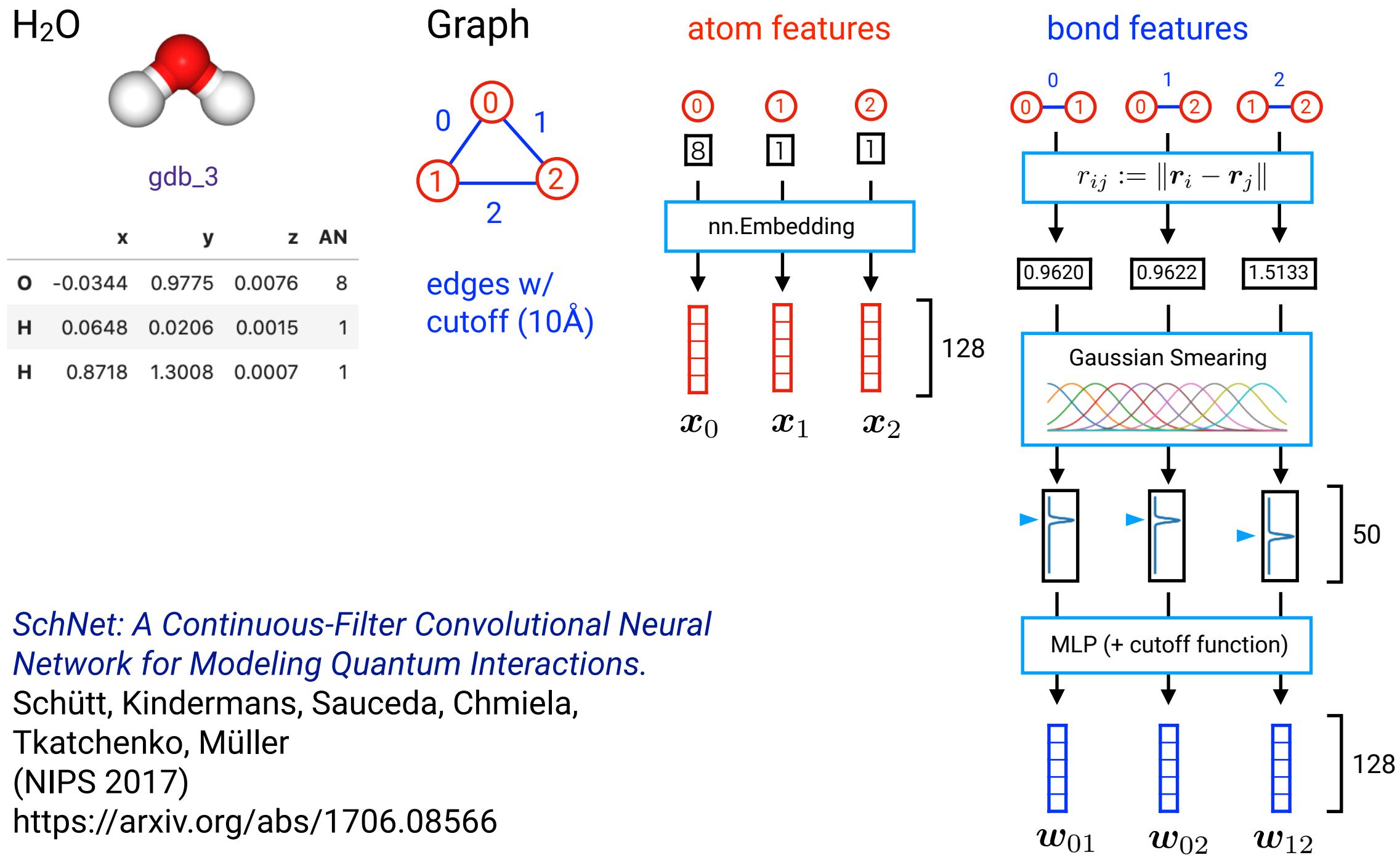
Schütt, Kindermans, Sauceda, Chmiela,

Tkatchenko, Müller

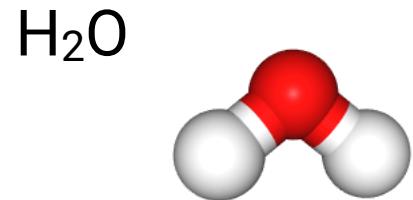
(NIPS 2017)

<https://arxiv.org/abs/1706.08566>

SchNet (Schütt et al, 2017): 幾何的GNNの先駆的Standard

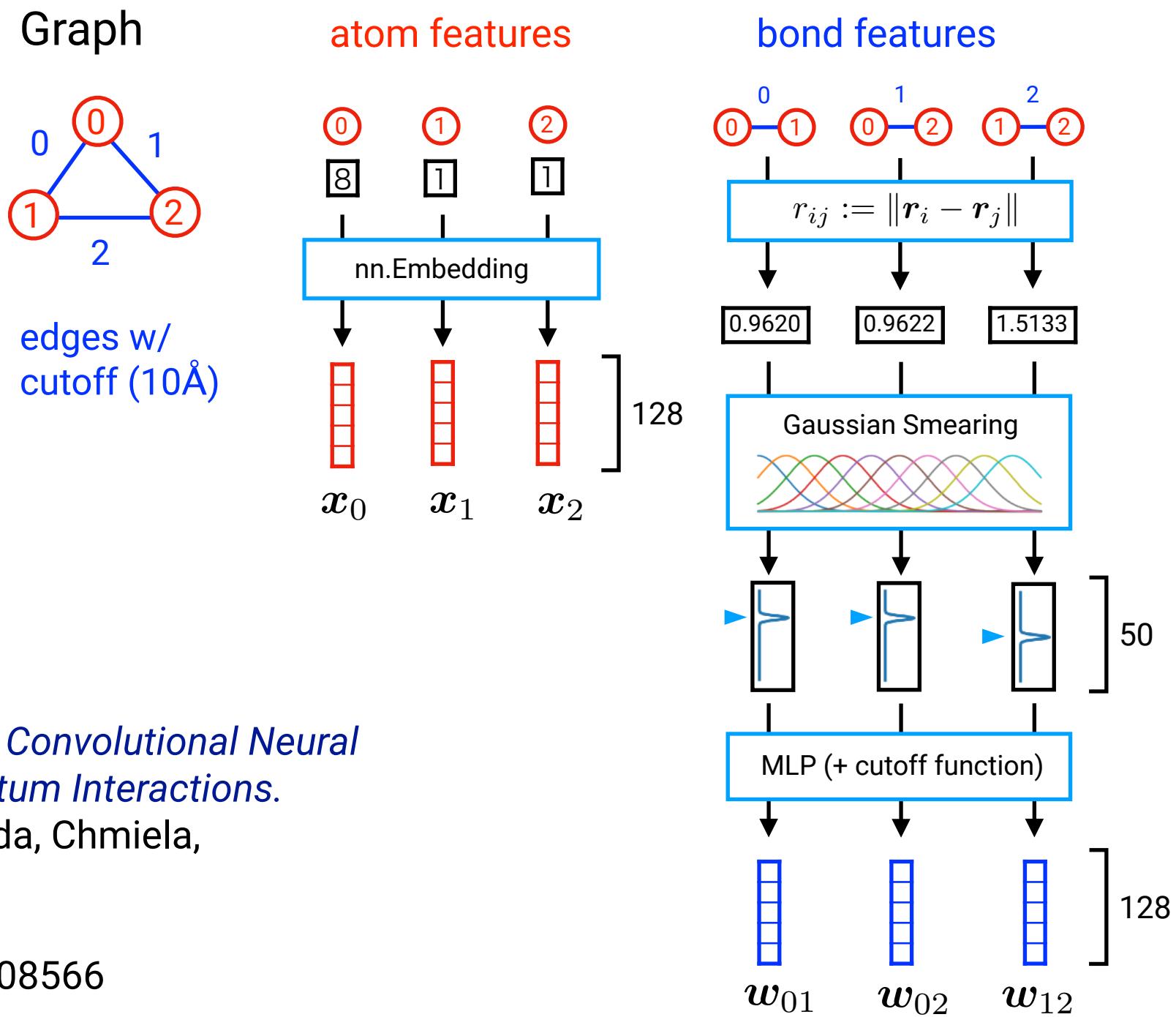


SchNet (Schütt et al, 2017): 幾何的GNNの先駆的Standard



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

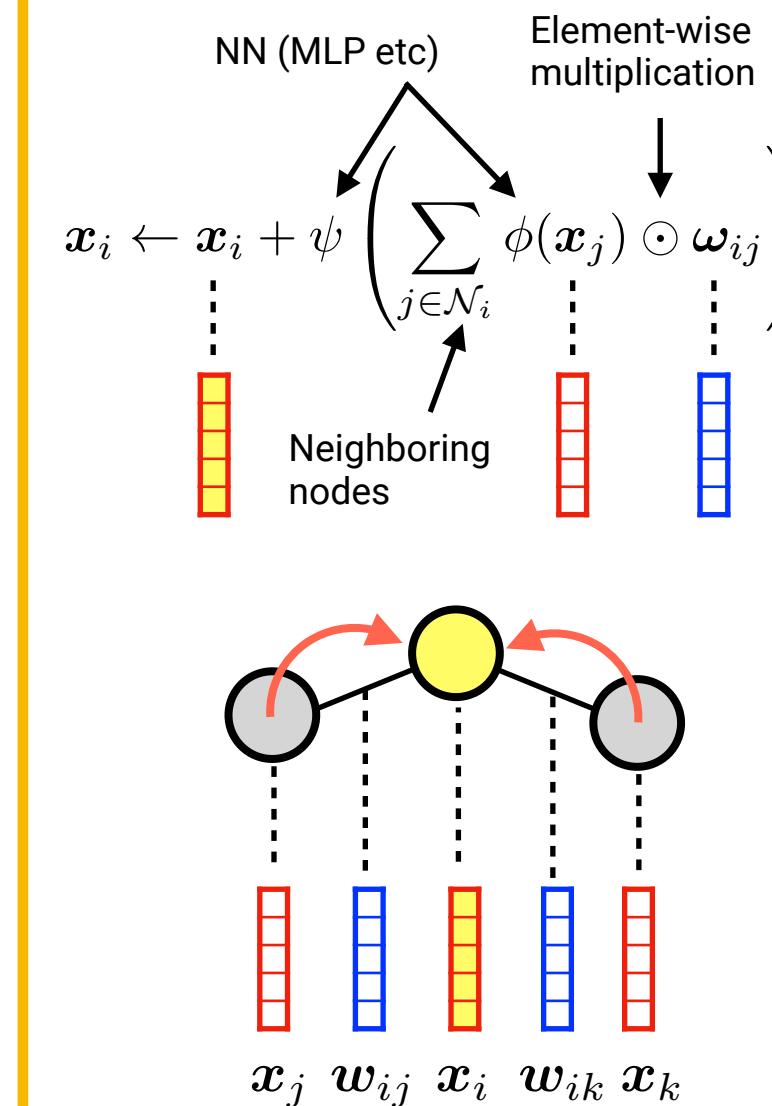


SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions.

Schütt, Kindermans, Sauceda, Chmiela,
Tkatchenko, Müller
(NIPS 2017)

<https://arxiv.org/abs/1706.08566>

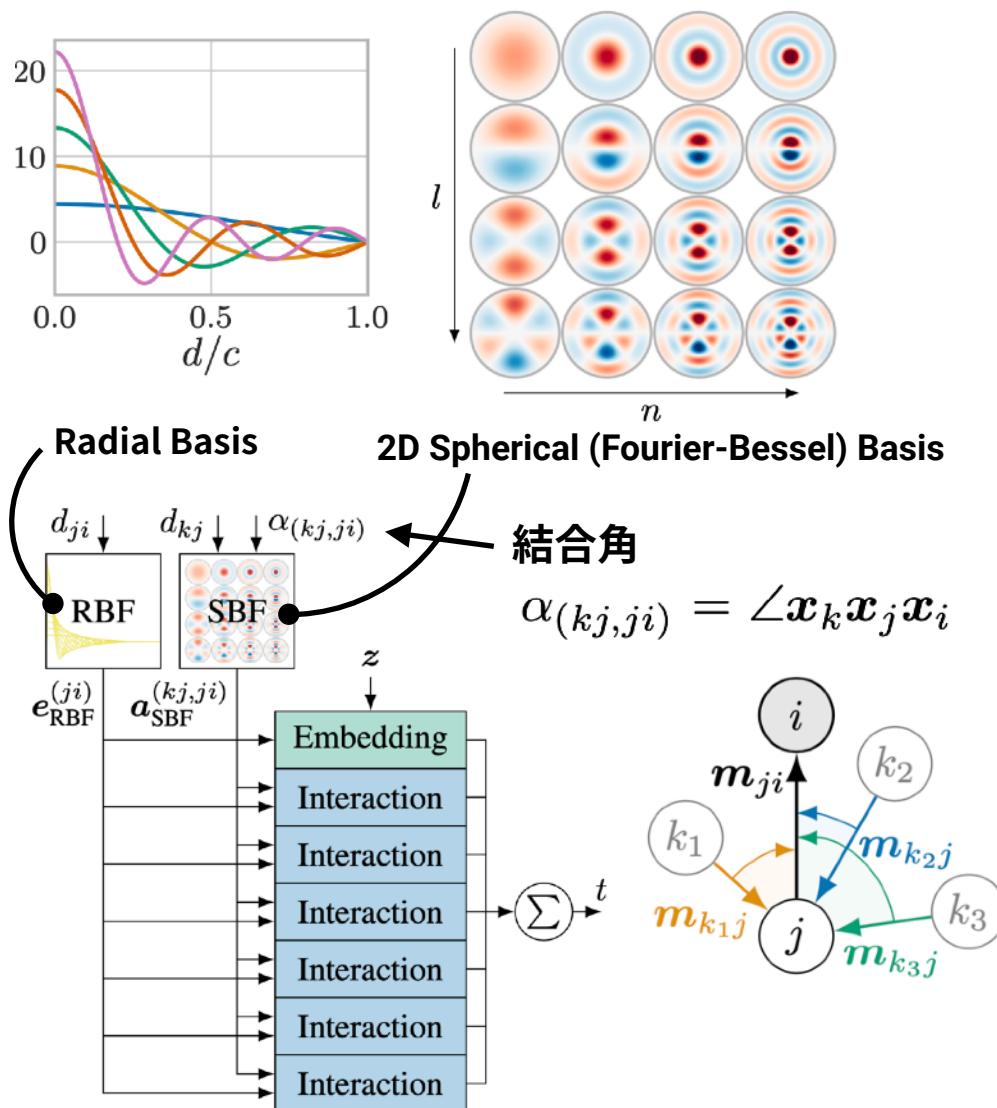
Message Passing with residual connections



分子に特化した帰納バイアスによる多様な幾何的GNN

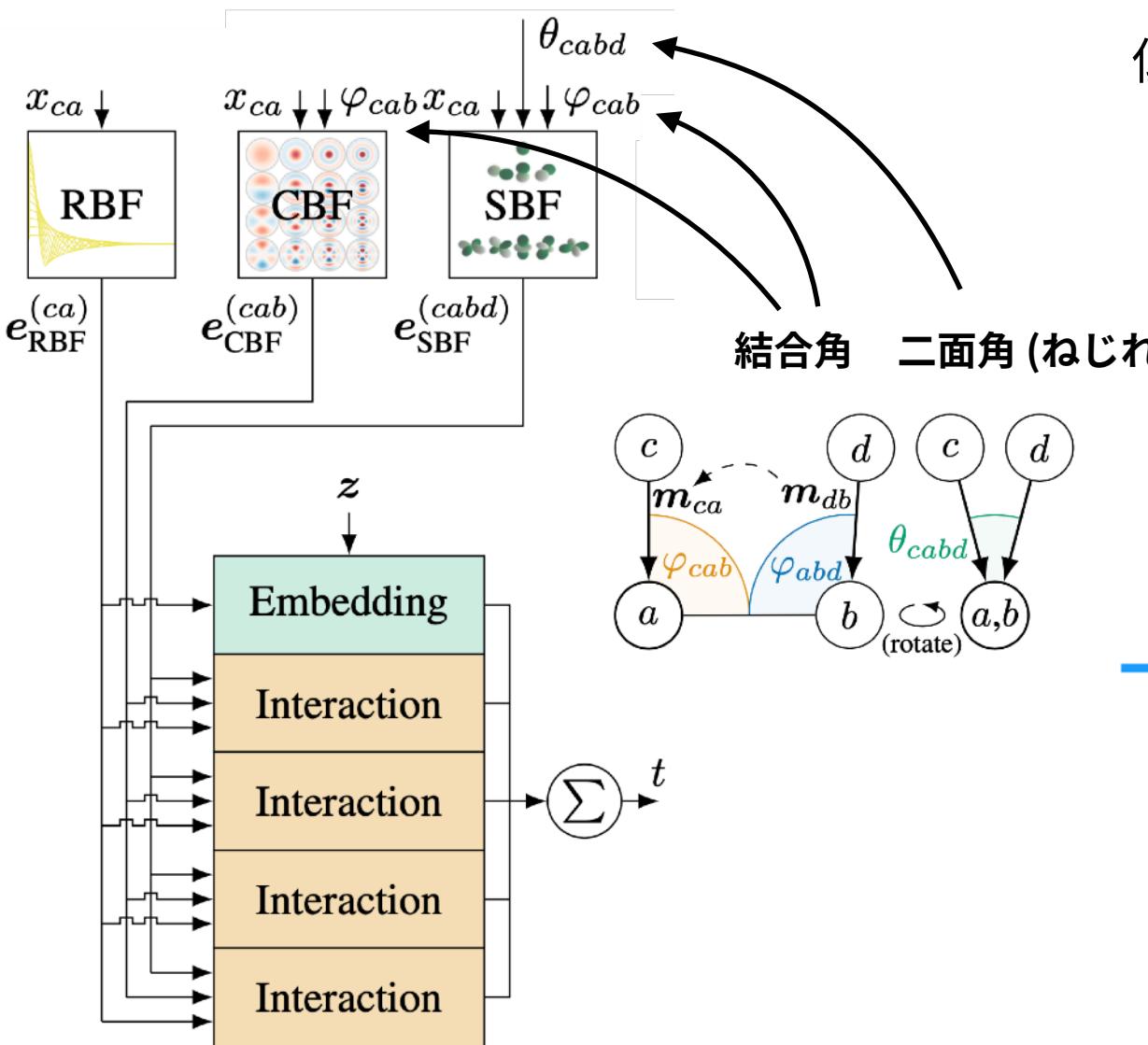
DimeNet++

Klicpera et al (NeurIPS WS2022)
<https://arxiv.org/abs/2011.14115>



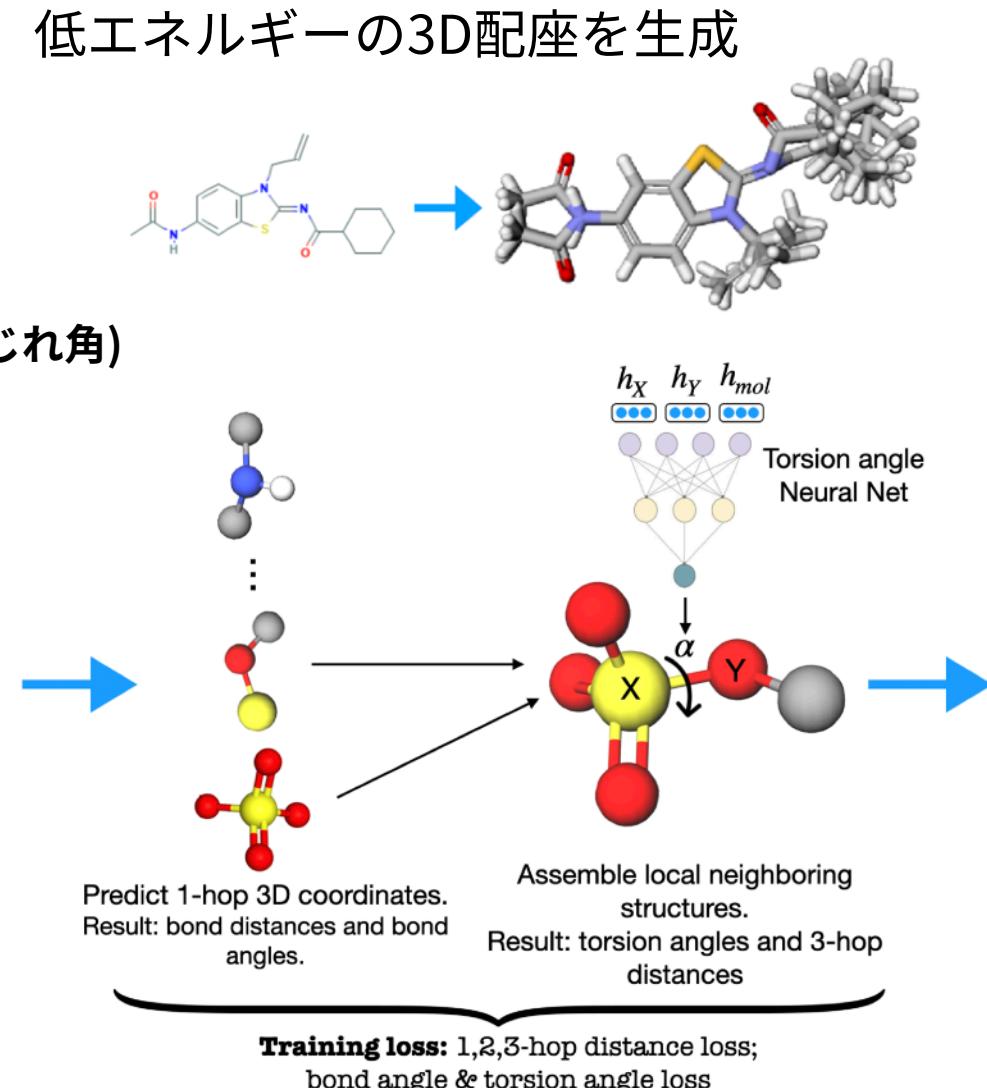
GemNet

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



GeoMol

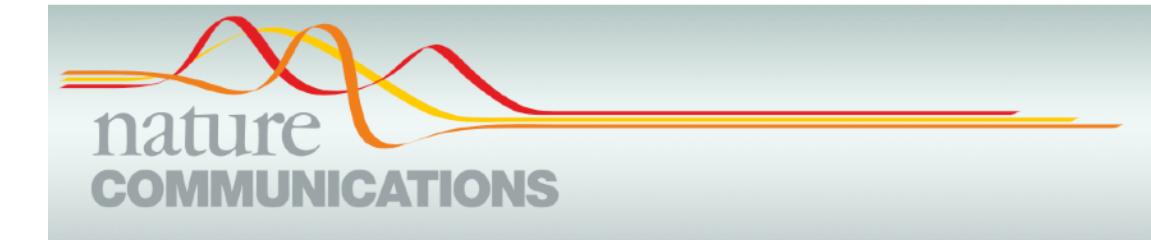
Ganea et al (NeurIPS2021)
<https://arxiv.org/abs/2106.07802>



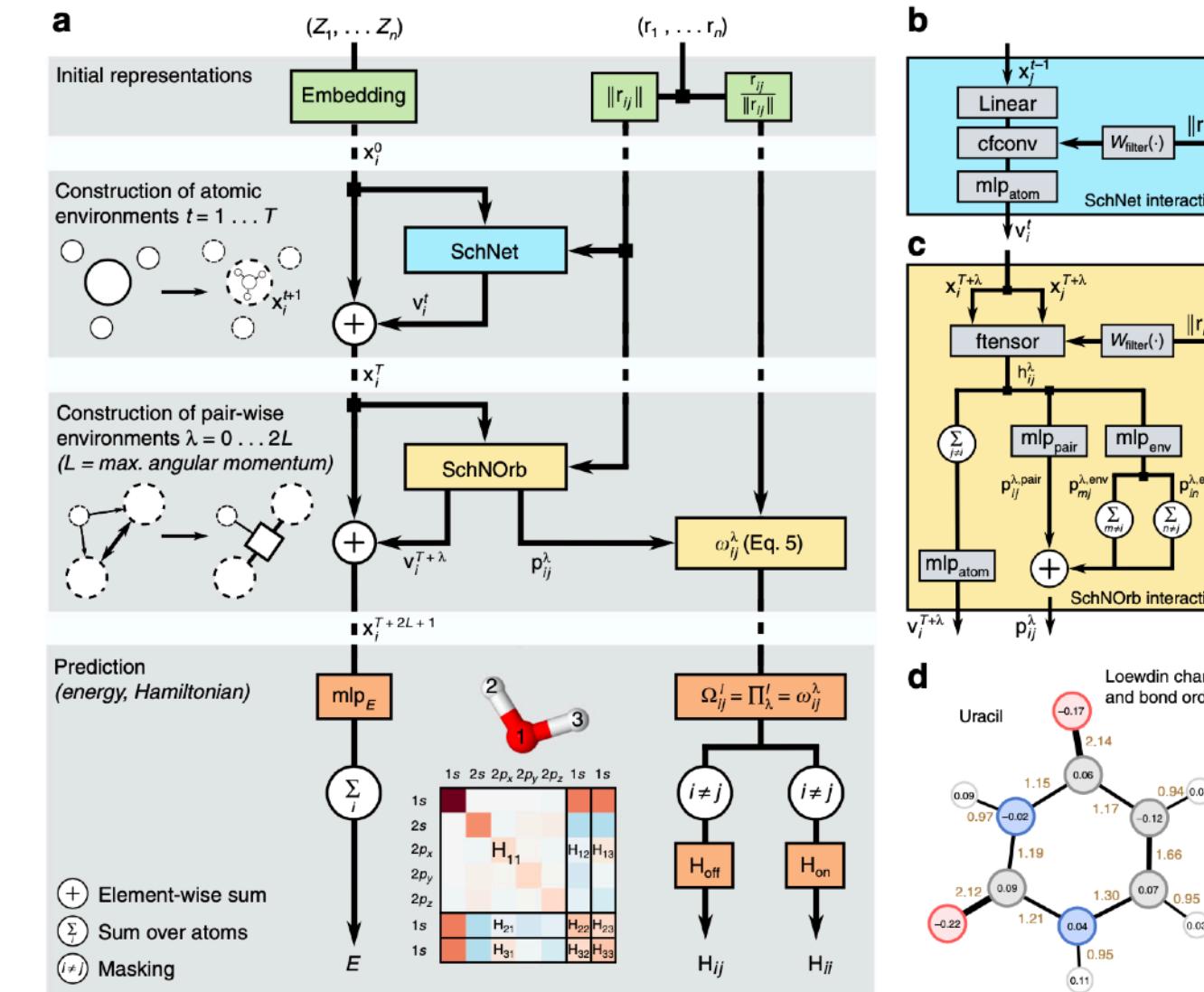
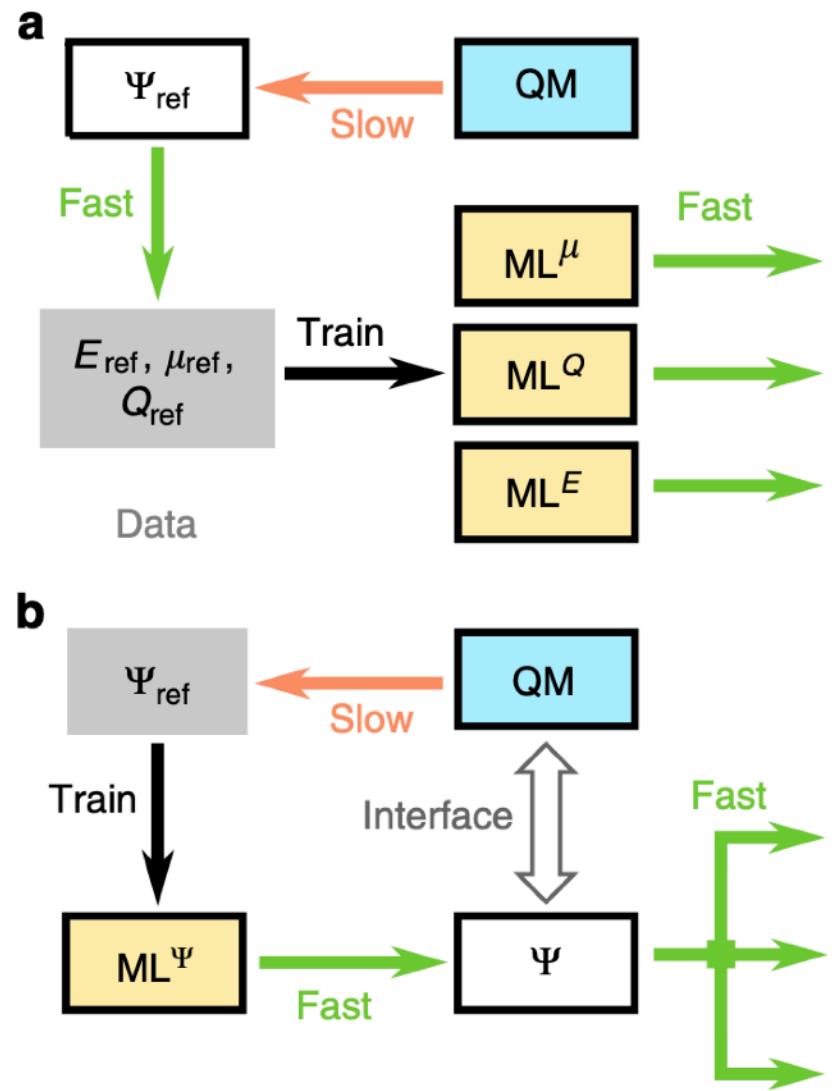
SchNOOrb: 波動関数(ハミルトニアン)自体を機械学習

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)



機械學習 × 量子化学計算

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



Learn to Simulate : ダイナミクスも同様に扱える !

DeepMind > Research > Learning to Simulate Complex Physics with Graph Networks

PUBLICATIONS

SHARE

PUBLICATION LINKS

DOWNLOAD

VIEW PUBLICATION

DATASETS & CODE

VIDEO SITE

→ VIEW OPEN SOURCE

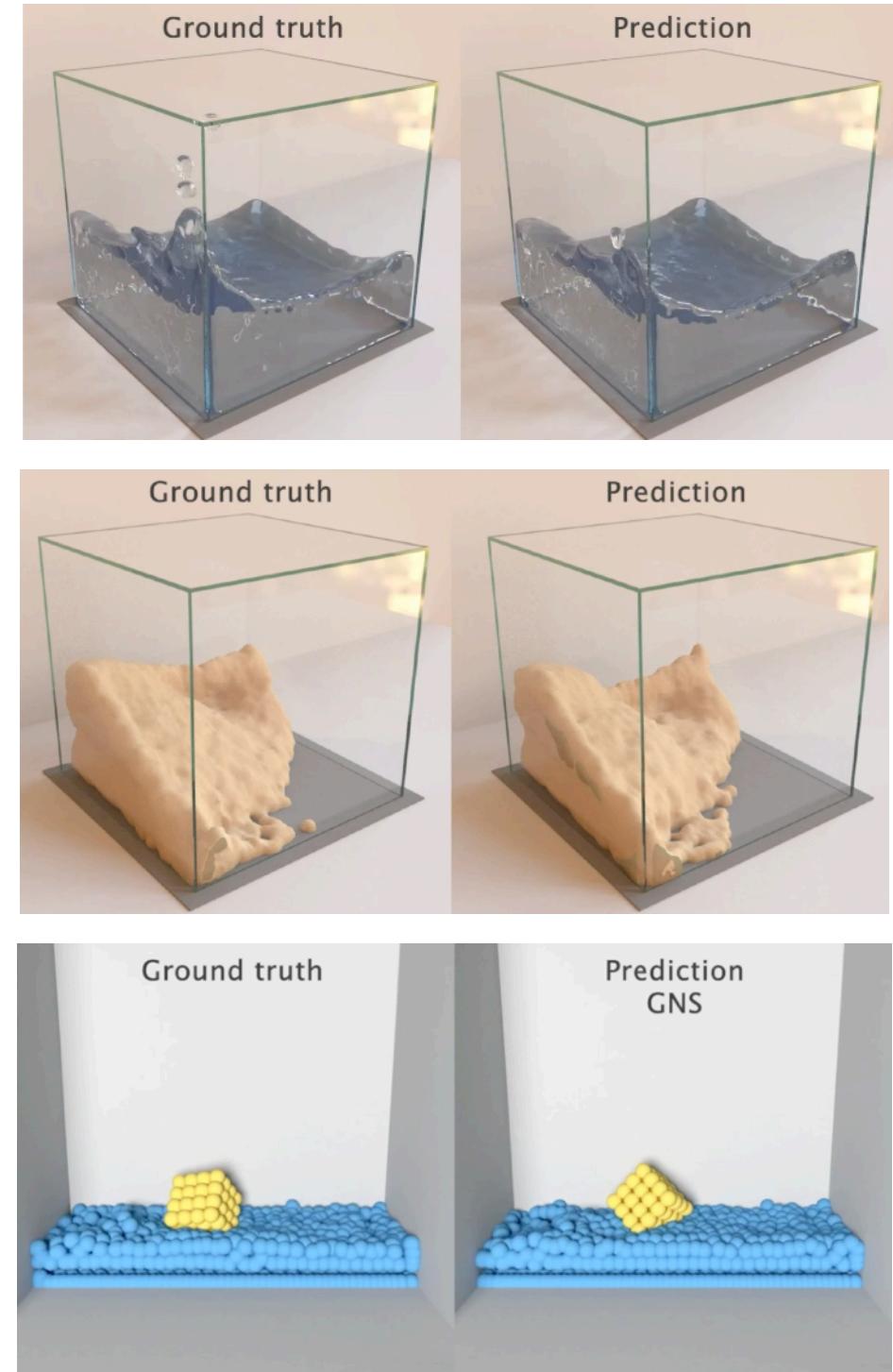
PUBLICATION ICML

Learning to Simulate Complex Physics with Graph Networks

Abstract

Here we present a machine learning framework and model implementation that can learn to simulate a wide variety of challenging physical domains, involving fluids, rigid solids, and deformable materials interacting with one another. Our framework—which we term “Graph Network-based Simulators” (GNS)—represents the state of a physical system with particles, expressed as nodes in a graph, and computes dynamics via learned message-passing. Our results show that our model can generalize from single-timestep predictions with thousands of particles during training, to different initial conditions, thousands of timesteps, and at least an order of magnitude more particles at test time. Our model was robust to hyperparameter choices across various evaluation metrics: the main determinants of long-term performance were the number of message-passing steps, and mitigating the accumulation of error by corrupting the training data with noise. Our GNS framework advances the state-of-the-art in learned physical simulation, and holds promise for solving a wide range of complex forward and inverse problems.

Datasets and example model and training code available.



Learn to Simulate : ダイナミクスも同様に扱える !

DeepMind > Research > Learning to Simulate Complex Physics with Graph Networks

PUBLICATIONS

SHARE

PUBLICATION LINKS

DOWNLOAD

VIEW PUBLICATION

DATASETS & CODE

VIDEO SITE

→ VIEW OPEN SOURCE

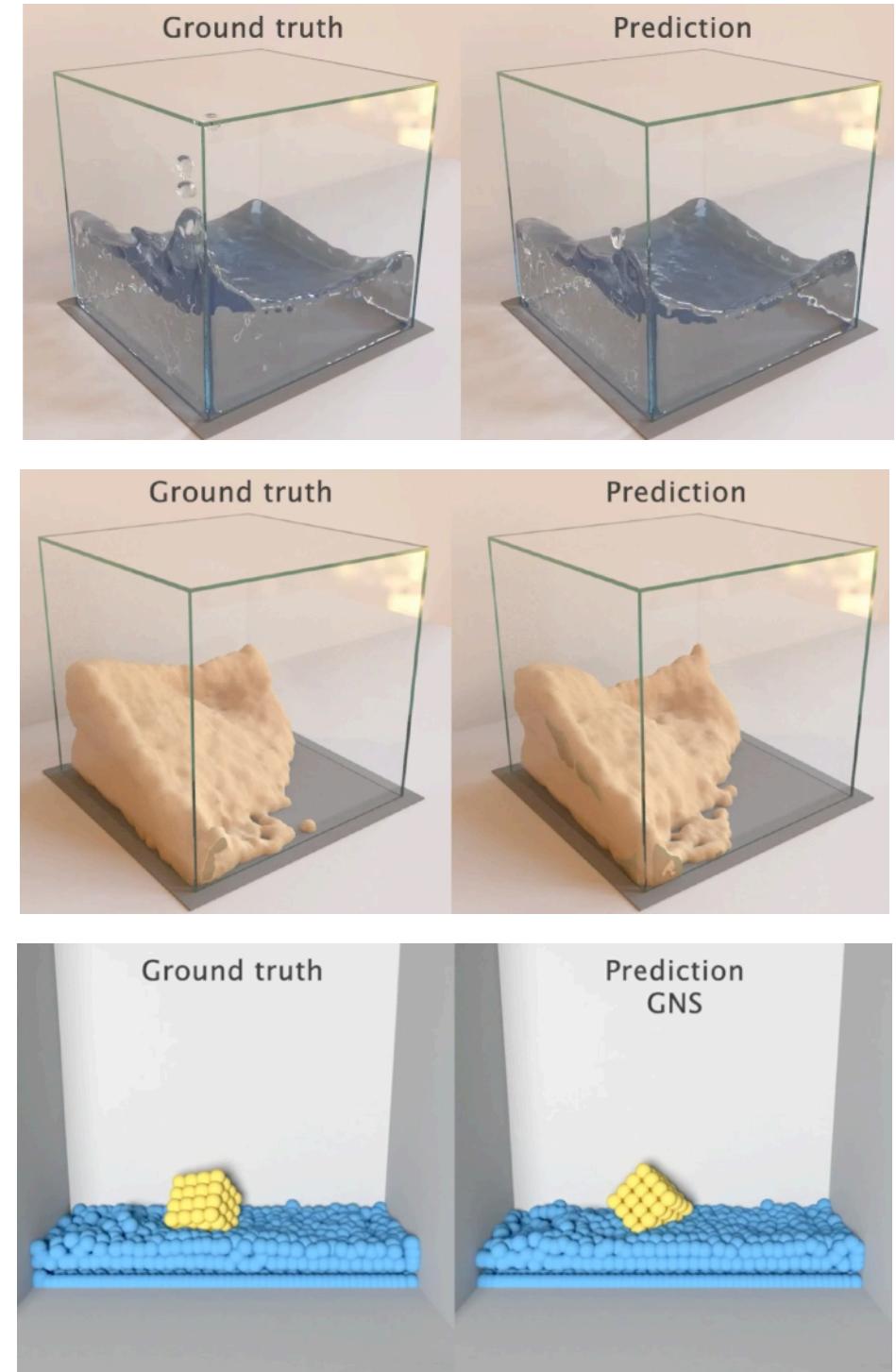
PUBLICATION ICML

Learning to Simulate Complex Physics with Graph Networks

Abstract

Here we present a machine learning framework and model implementation that can learn to simulate a wide variety of challenging physical domains, involving fluids, rigid solids, and deformable materials interacting with one another. Our framework—which we term “Graph Network-based Simulators” (GNS)—represents the state of a physical system with particles, expressed as nodes in a graph, and computes dynamics via learned message-passing. Our results show that our model can generalize from single-timestep predictions with thousands of particles during training, to different initial conditions, thousands of timesteps, and at least an order of magnitude more particles at test time. Our model was robust to hyperparameter choices across various evaluation metrics: the main determinants of long-term performance were the number of message-passing steps, and mitigating the accumulation of error by corrupting the training data with noise. Our GNS framework advances the state-of-the-art in learned physical simulation, and holds promise for solving a wide range of complex forward and inverse problems.

Datasets and example model and training code available.



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)



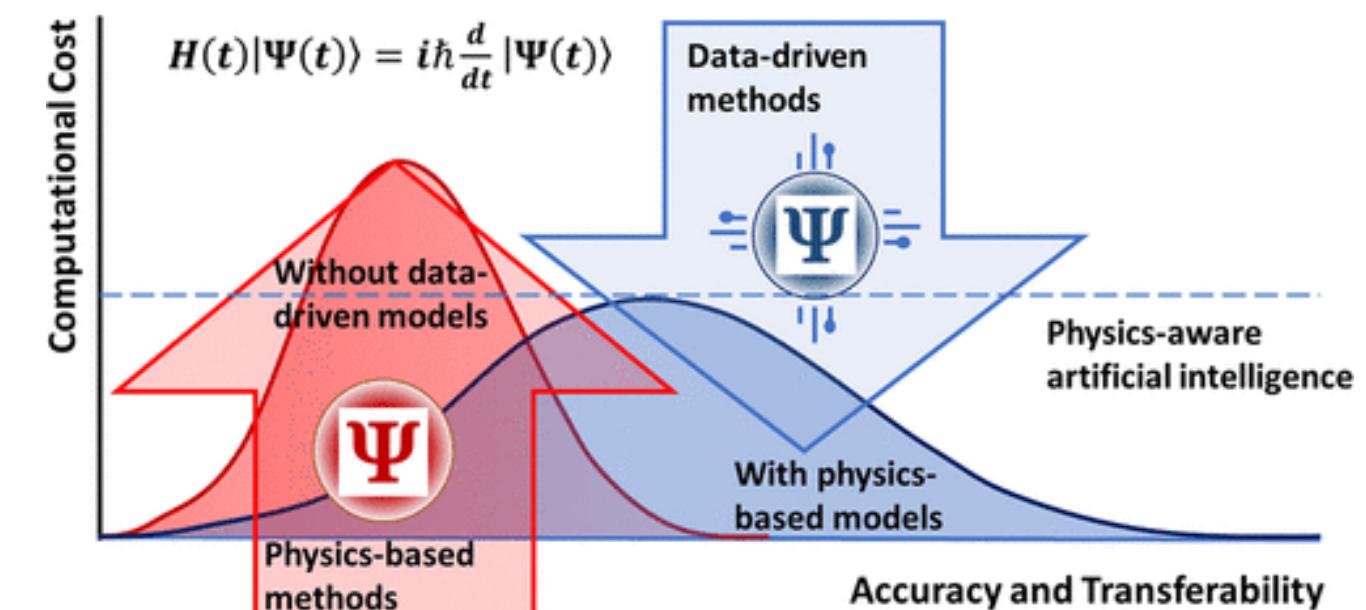
pubs.acs.org/accounts

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×知識処理・論理推論・プランニング

メモリからの適応的読み出しを含む手続き的・記号的操も学習で扱えるようになってきた！

Neural Abstract Machines & Program Induction

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

- **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)
- :



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



知識ベースに収集された明示的な化学的な知識も融合していくか？

Reaxys®

SCI-FINDER®
A CAS SOLUTION

CHEMATICAA

**すごく面白い技術課題が山のようになり技術屋にとって
既にものすごく楽しい！**

**すごく面白い技術課題が山のようになり技術屋にとって
既にものすごく楽しい！**

**…が、最初に念押しした「本当のゴール」を君はまさか
忘れてしまってはいないだろうな？**

本当の戦いは以上のイケてる技術を武器にここから始まる！

Hello

World

ダークサイドへようこそ：こんにちは、世界！

ダークサイドへようこそ：こんにちは、世界！

- ✓ 華々しい成果は今のところ主に「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・
大規模なオープンデータが利用できる・etc

ダークサイドへようこそ：こんにちは、世界！

- ✓ 華々しい成果は今のところ主に「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・
大規模なオープンデータが利用できる・etc
- ✓ 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ねえんだって…

ダークサイドへようこそ：こんにちは、世界！

- ✓ 華々しい成果は今のところ主に「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・
 - 大規模なオープンデータが利用できる・etc
- ✓ 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ねえんだって…
 - 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)

ダークサイドへようこそ：こんにちは、世界！

- ✓ 華々しい成果は今のところ主に「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・
 - 大規模なオープンデータが利用できる・etc
- ✓ 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ねえんだって…
 - 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

ダークサイドへようこそ：こんにちは、世界！

- ✓ 華々しい成果は今のところ主に「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・
大規模なオープンデータが利用できる・etc
- ✓ 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ねえんだって…
 - 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

ダークサイドへようこそ：こんにちは、世界！

- ✓ 華々しい成果は今のところ主に「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・
大規模なオープンデータが利用できる・etc
- ✓ 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ねえんだって…
 - 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

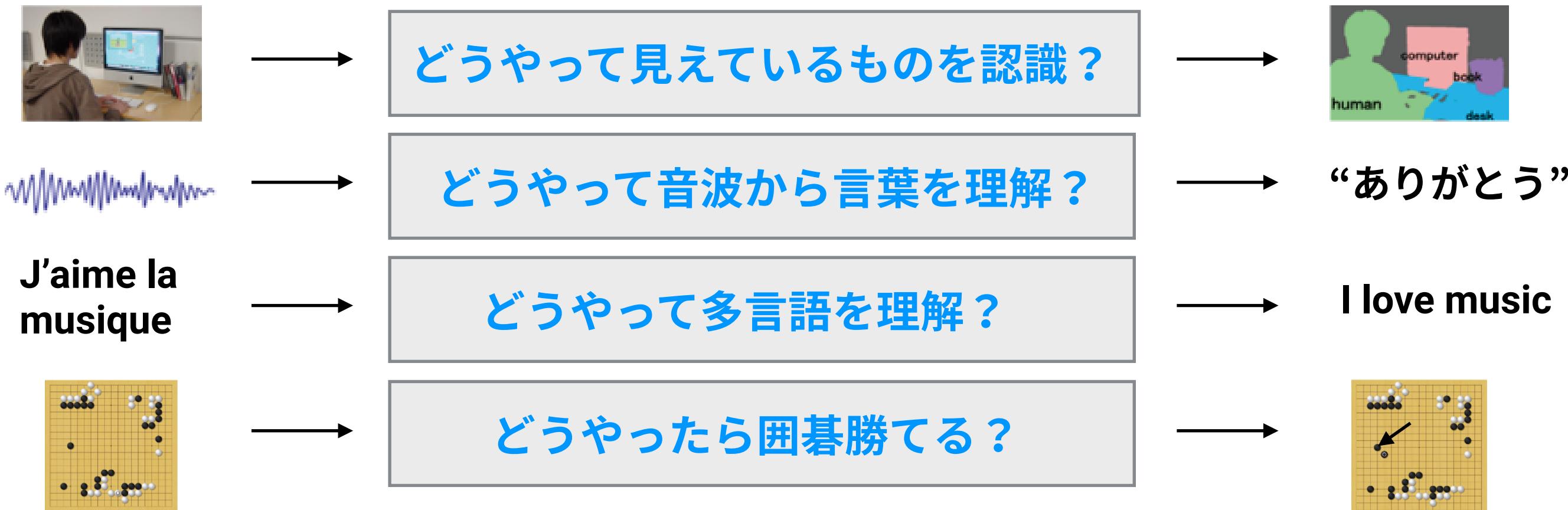
ダークサイドへようこそ：こんにちは、世界！

- ✓ 華々しい成果は今のところ主に「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・
 - 大規模なオープンデータが利用できる・etc
- ✓ 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ねえんだって…
 - 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
 - 何か学習させるとときに「良い例題や良い演習問題」って本当に大事ですよね！！

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でしょうか？



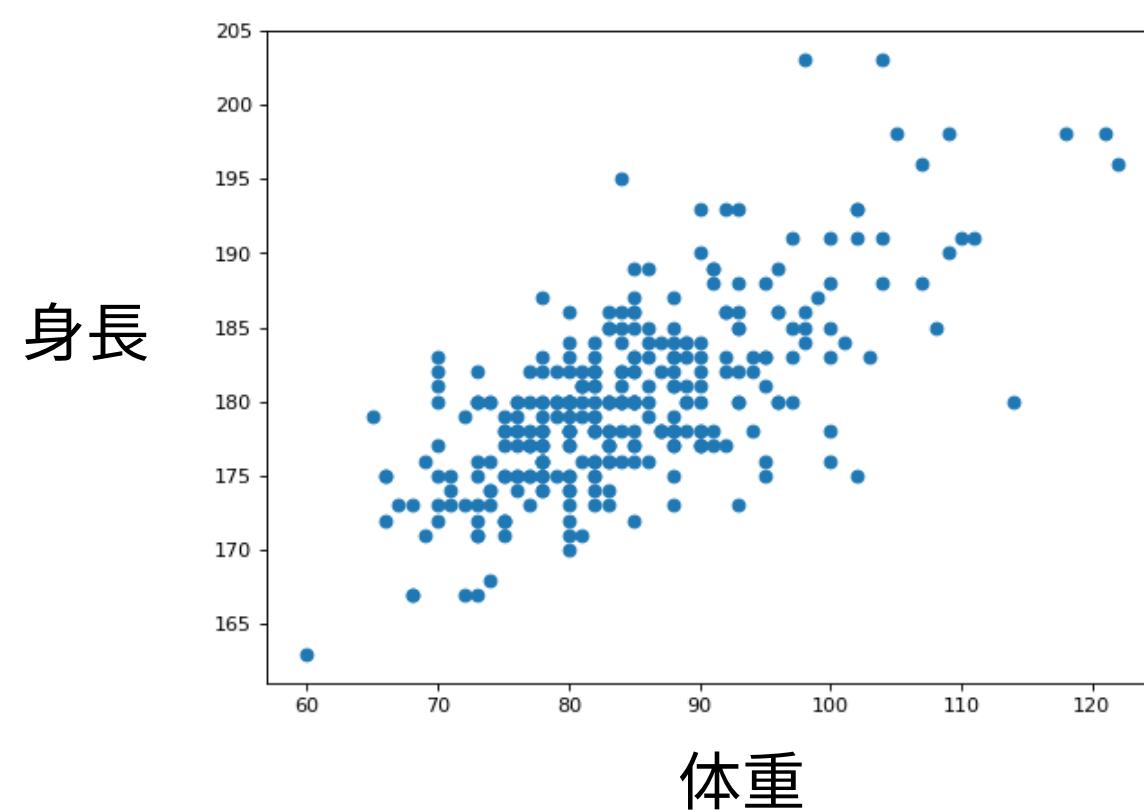
因果の理解には実験研究(介入研究)が必要不可欠

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

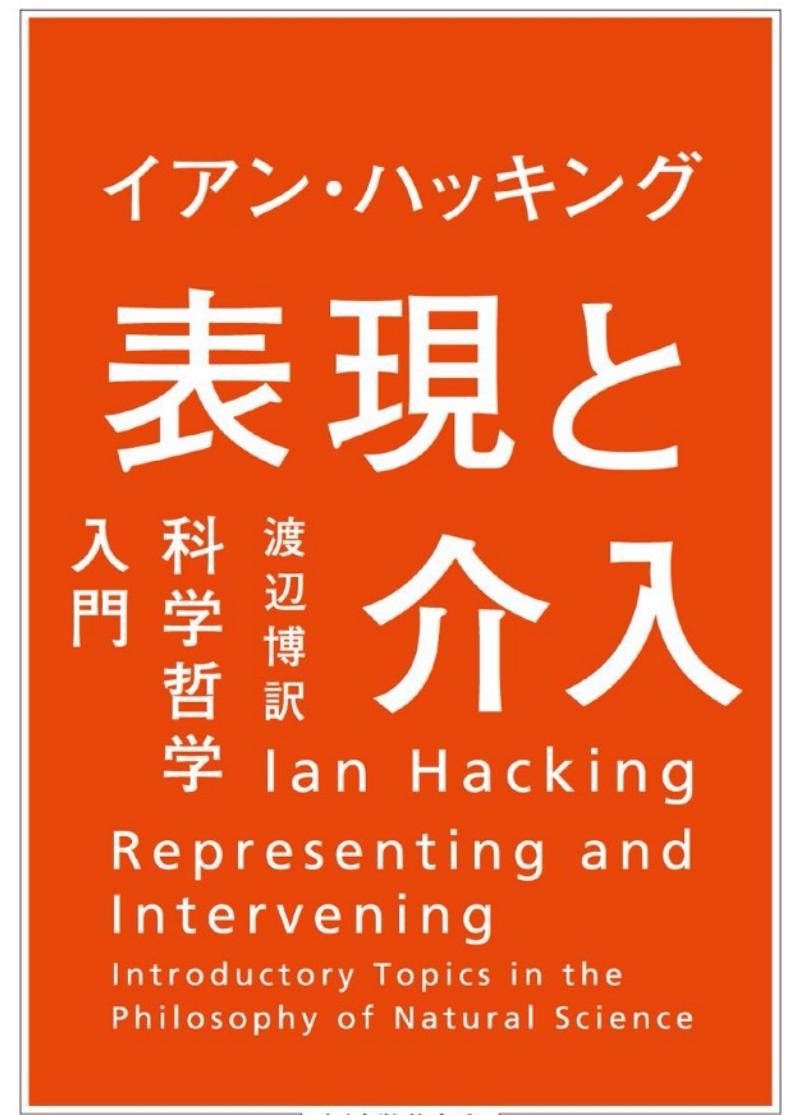
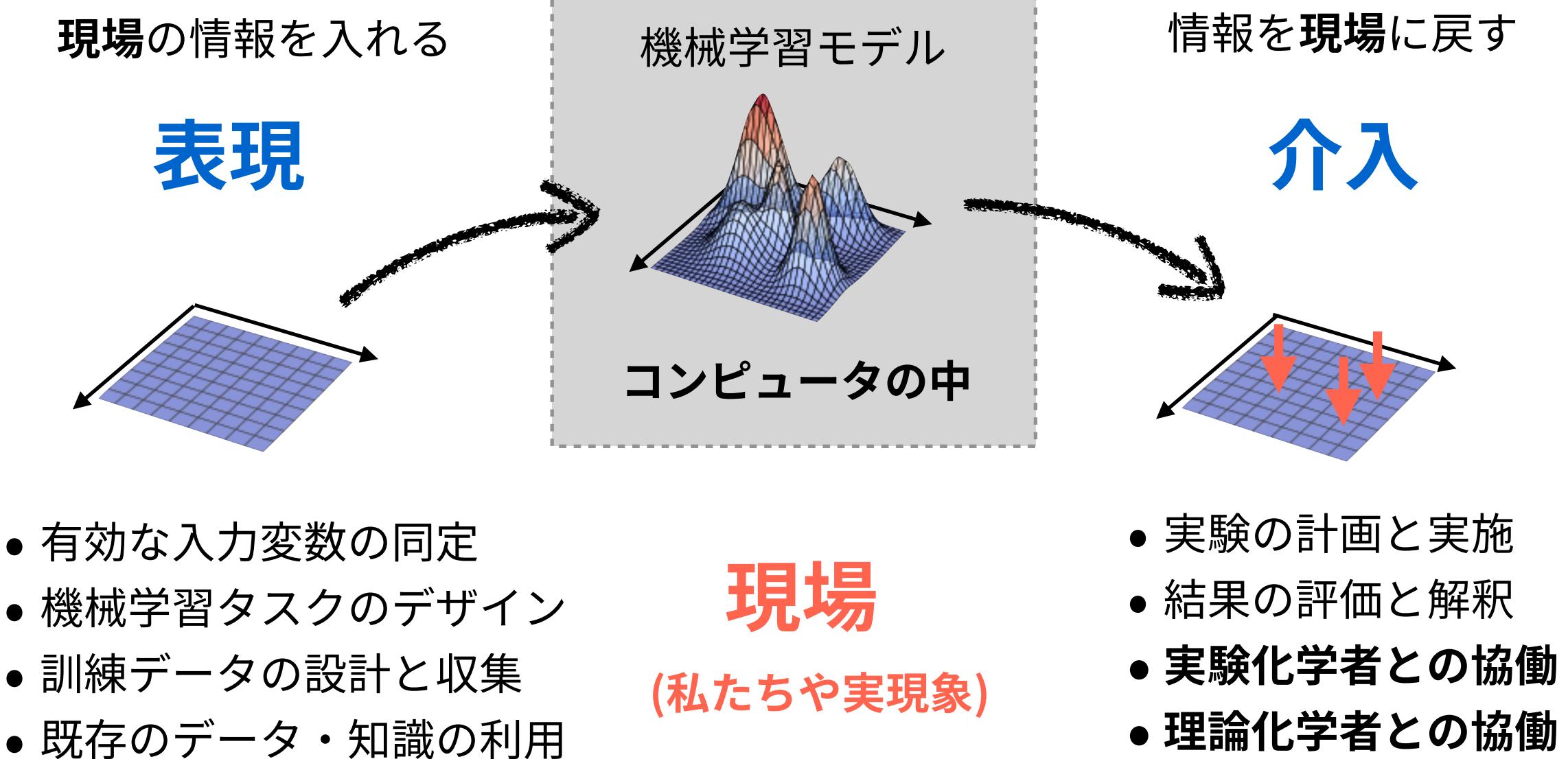
応用統計学の基本のキ

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！！}

予測から理解・発見へ：頭でっかちを脱し、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 僕



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！

- **解釈性**：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！

- **解釈性**：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。
- **時間性**：**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！

- **解釈性**：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。
- **時間性**：**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！

- **解釈性**：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。
- **時間性**：**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。
- **選択バイアス**：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。与えられたデータの傾向を捉える機械学習の予測も同様にそ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！

- **解釈性**：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。
- **時間性**：**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。
- **選択バイアス**：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。与えられたデータの傾向を捉える機械学習の予測も同様にその偏りから逃れられない。
- **因果性の理解**：「**因果性**」は直接観測できない。人間がアクセスできるのは「相関」だけ！

機械学習から機械発見へ

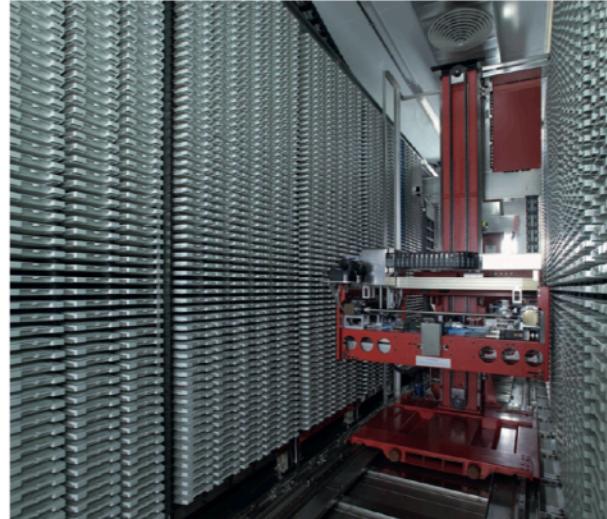
実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



Science 363 (2019)



Nature 583 (2020)



Nature Reviews Drug Discovery 17 (2018)



機械学習から機械発見へ

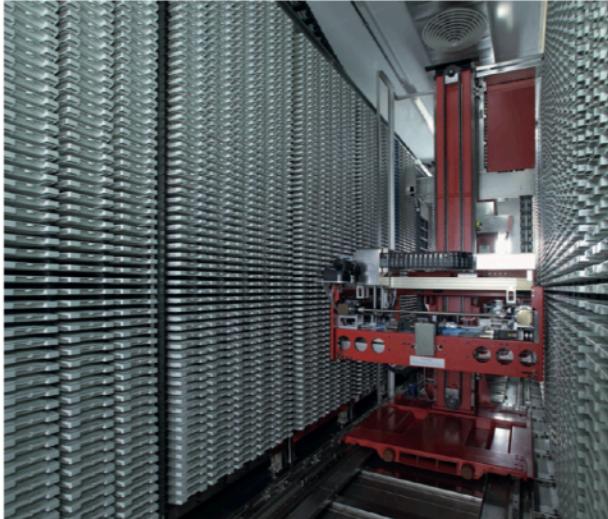
実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



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 + Negative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

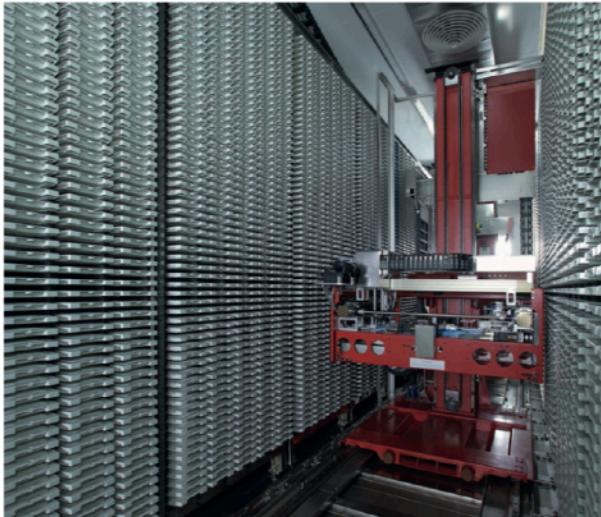
実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



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 + Negative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…
- 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

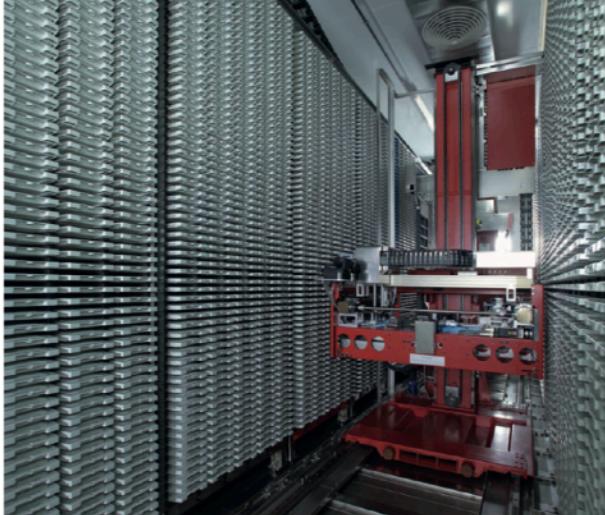
実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



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 + Negative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…
- 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**
- **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

まとめ：機械学習を自然現象の理解・発見に活用するとは？

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見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

まとめ：機械学習を自然現象の理解・発見に活用するとは？

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見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

- 新たな枠組み「見本例によるプログラミング」を「どこにどう使うか」のセンスが問われる。
「明示的な関係はよく分からないが入出力見本データは取れる」部分問題を熟考すること。

まとめ：機械学習を自然現象の理解・発見に活用するとは？

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見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**

- 新たな枠組み「見本例によるプログラミング」を「どこにどう使うか」のセンスが問われる。「明示的な関係はよく分からないが入出力見本データは取れる」部分問題を熟考すること。
- 現状では「本質的にはデータが足りてない」場合がほとんどであり、専門家と機械学習屋が協働で分野の今までの知識や知見を生かし上手に「帰納バイアス」を設計する必要がある。

まとめ：機械学習を自然現象の理解・発見に活用するとは？

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見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**

- 新たな枠組み「見本例によるプログラミング」を「どこにどう使うか」のセンスが問われる。「明示的な関係はよく分からないが入出力見本データは取れる」部分問題を熟考すること。
- 現状では「本質的にはデータが足りてない」場合がほとんどであり、専門家と機械学習屋が協働で分野の今までの知識や知見を生かし上手に「帰納バイアス」を設計する必要がある。
- 「因果性」は直接観測できないので、「実際に実験によって確かめてみる」介入が不可欠。この検証ステップをどのようにデザイン・実現するかが非常に大切。

まとめ：機械学習を自然現象の理解・発見に活用するとは？

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見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**

- 新たな枠組み「見本例によるプログラミング」を「どこにどう使うか」のセンスが問われる。「明示的な関係はよく分からないが入出力見本データは取れる」部分問題を熟考すること。
- 現状では「本質的にはデータが足りてない」場合がほとんどであり、専門家と機械学習屋が協働で分野の今までの知識や知見を生かし上手に「帰納バイアス」を設計する必要がある。
- 「因果性」は直接観測できないので、「実際に実験によって確かめてみる」介入が不可欠。この検証ステップをどのようにデザイン・実現するかが非常に大切。
- 「大規模データが得られる設定では非常に強力な技術」なので近視眼的に今のところ手に入るデータだけで何とか場当たり的に頑張り続ける以上の中長期的なデータ獲得戦略が大事。

May the ML force be with you...

このスライドのPDFはこちらへ置いておきます → <https://itakigawa.github.io/news.html>

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」

