



SCIENCE AGORA 2021

データを予測に変える技術

機械学習で化粧に挑む!



龍川一学



たきがわ いちがく  
瀧川 一学

<https://itakigawa.github.io>

## 機械学習を研究している技術屋デス！

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

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



たきがわ いちがく  
瀧川 一学

<https://itakigawa.github.io>

# 機械学習を研究している技術屋デス！

- 香川県高松市生まれ

札幌 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)

京都 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)

札幌 規模知識処理研究室 准教授

2015～2015 JSTさきがけ (材料インフォマティクス)

京都 2019～ 北海道大学 化学反応創成研究拠点(ICReDD)

2019～ 理化学研究所 革新知能統合研究センター(AIP)

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

1. 機械学習は「データを予測に変える」
2. 機械学習は「新しい（そしてめっちゃ雑な！）コンピュータプログラムの作り方」

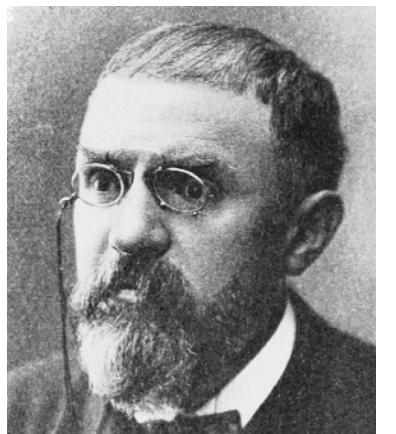
3. 現在の機械学習モデルはアホみたいにデータを食う…

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

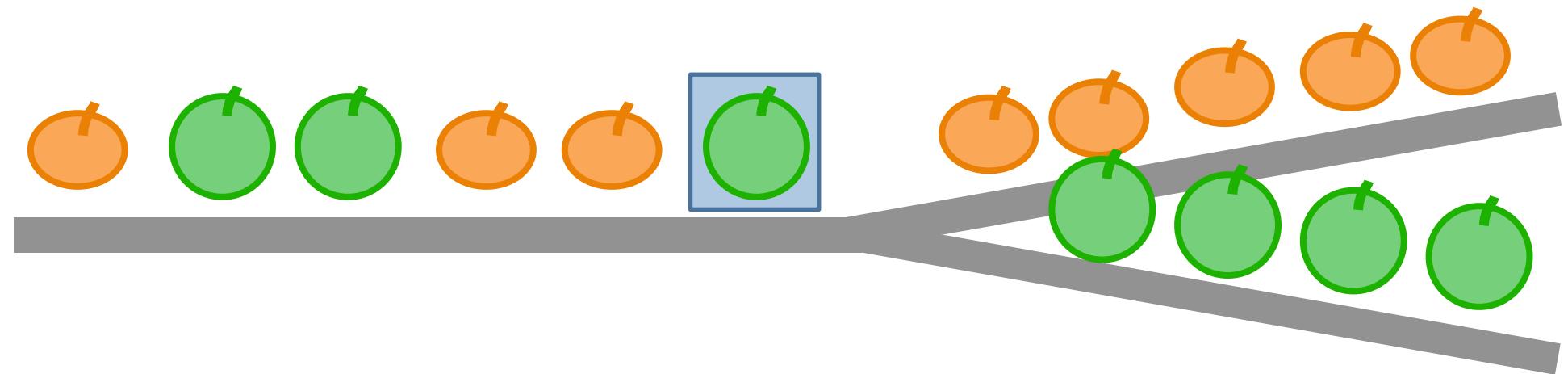
4. <sup>シン</sup>機械学習×化学の真の問題：機械学習から機械発見へ、予測から理解・発見へ  
事件はコンピュータ(機械学習)の中で起きてるんじゃない、現場で起きているんだ！ by 僕

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。

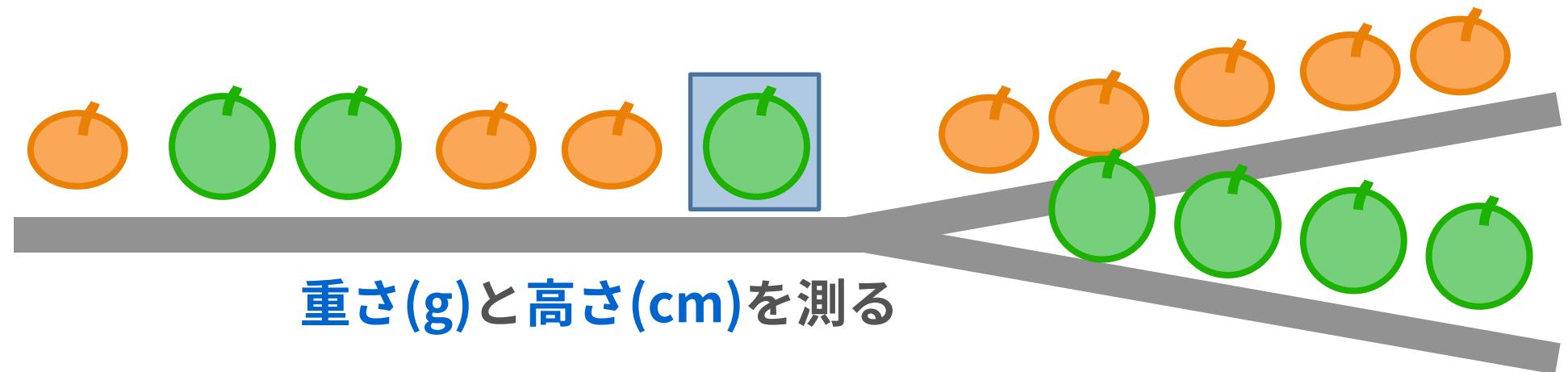
アンリ・ポアンカレ「科学と仮説」



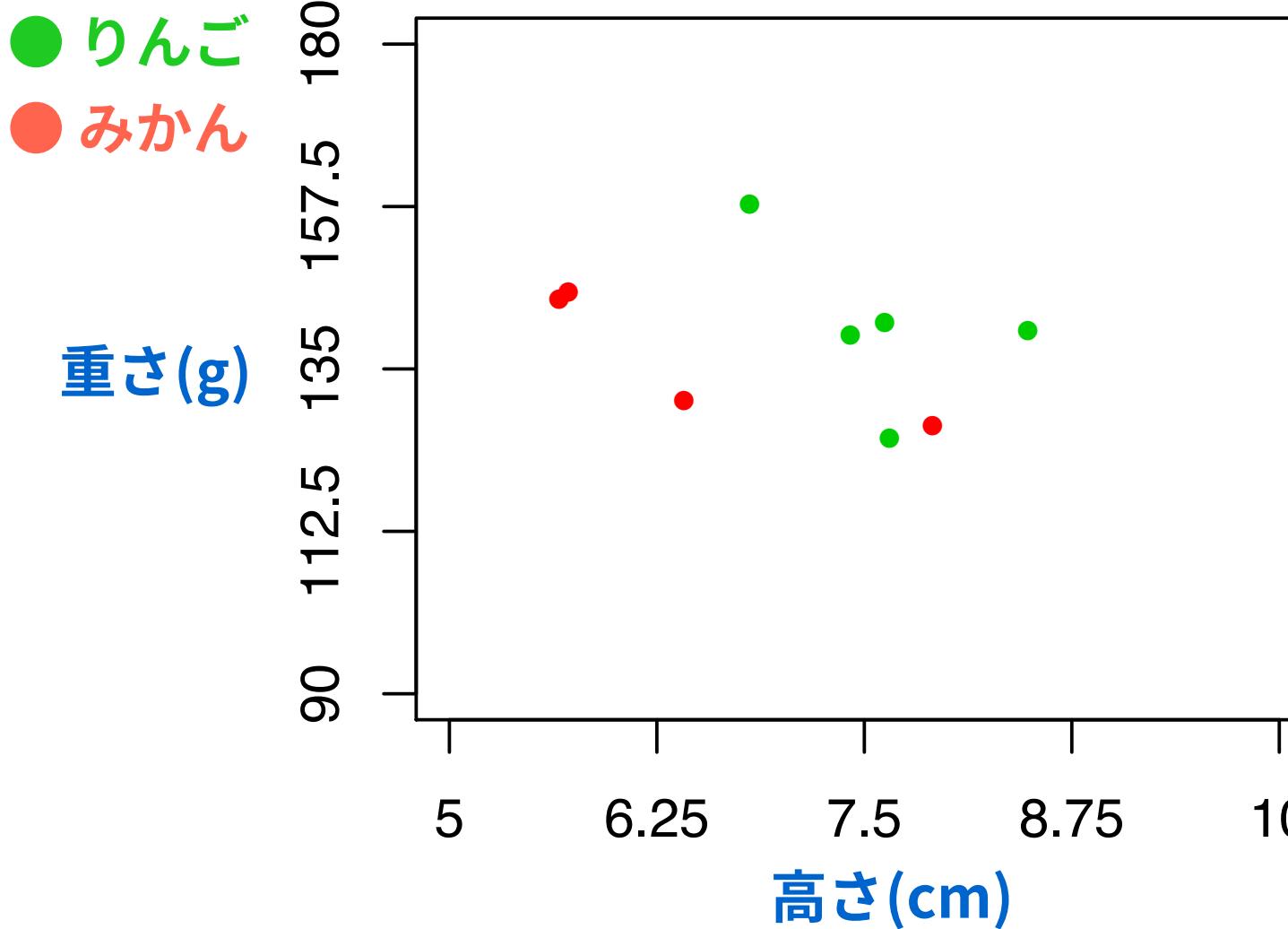
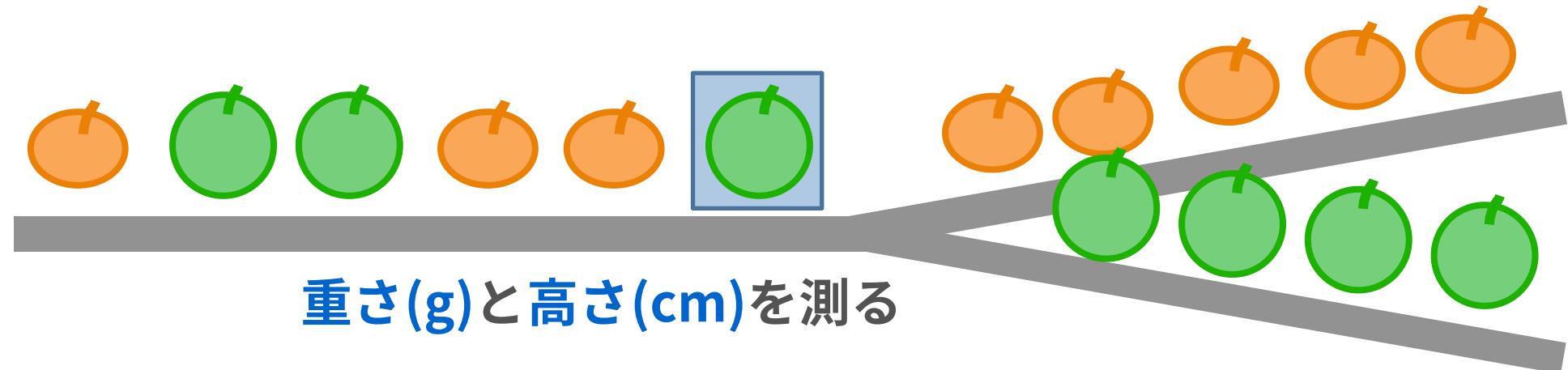
# 機械学習は「データを予測に変える」



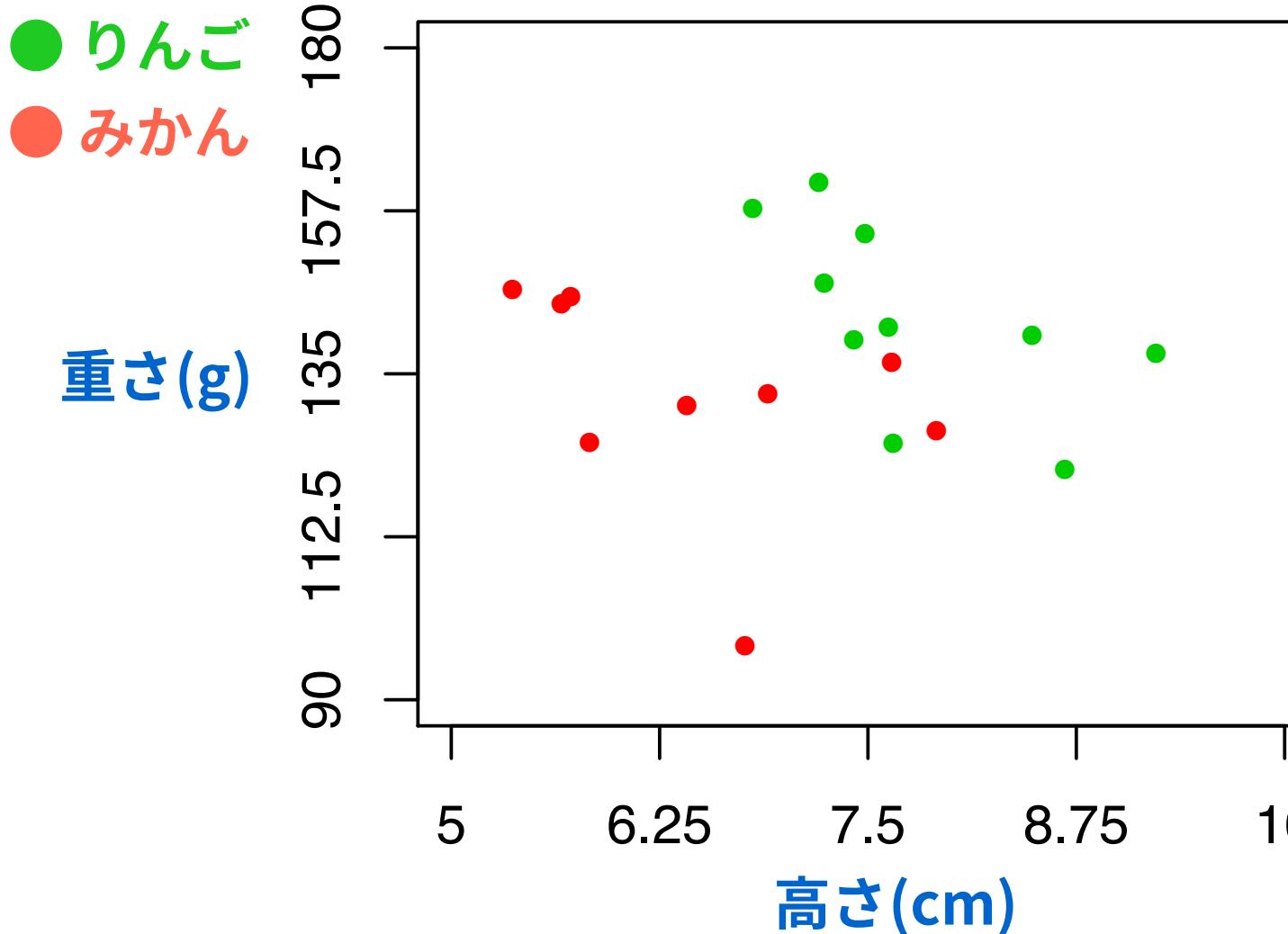
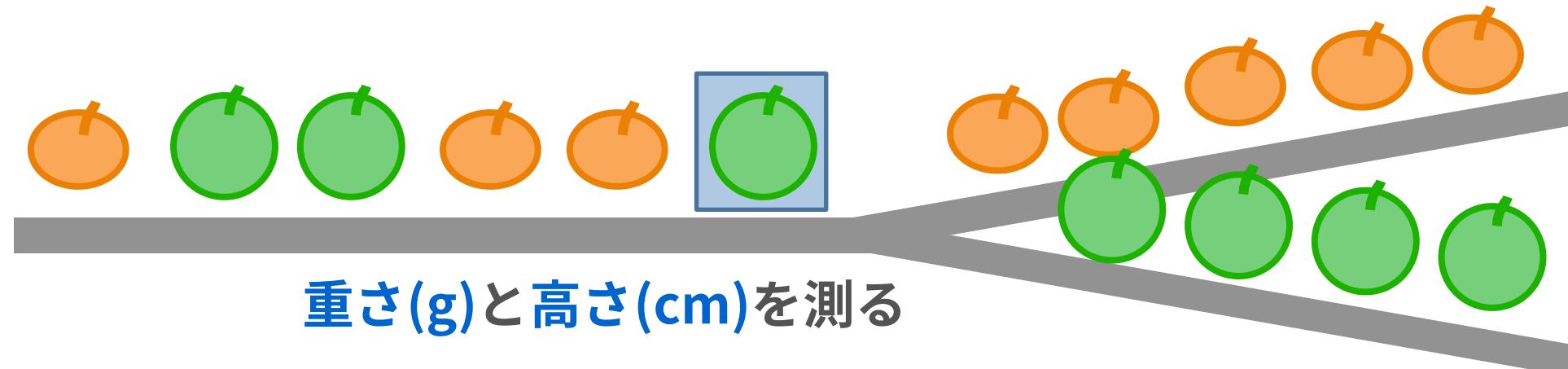
# 機械学習は「データを予測に変える」



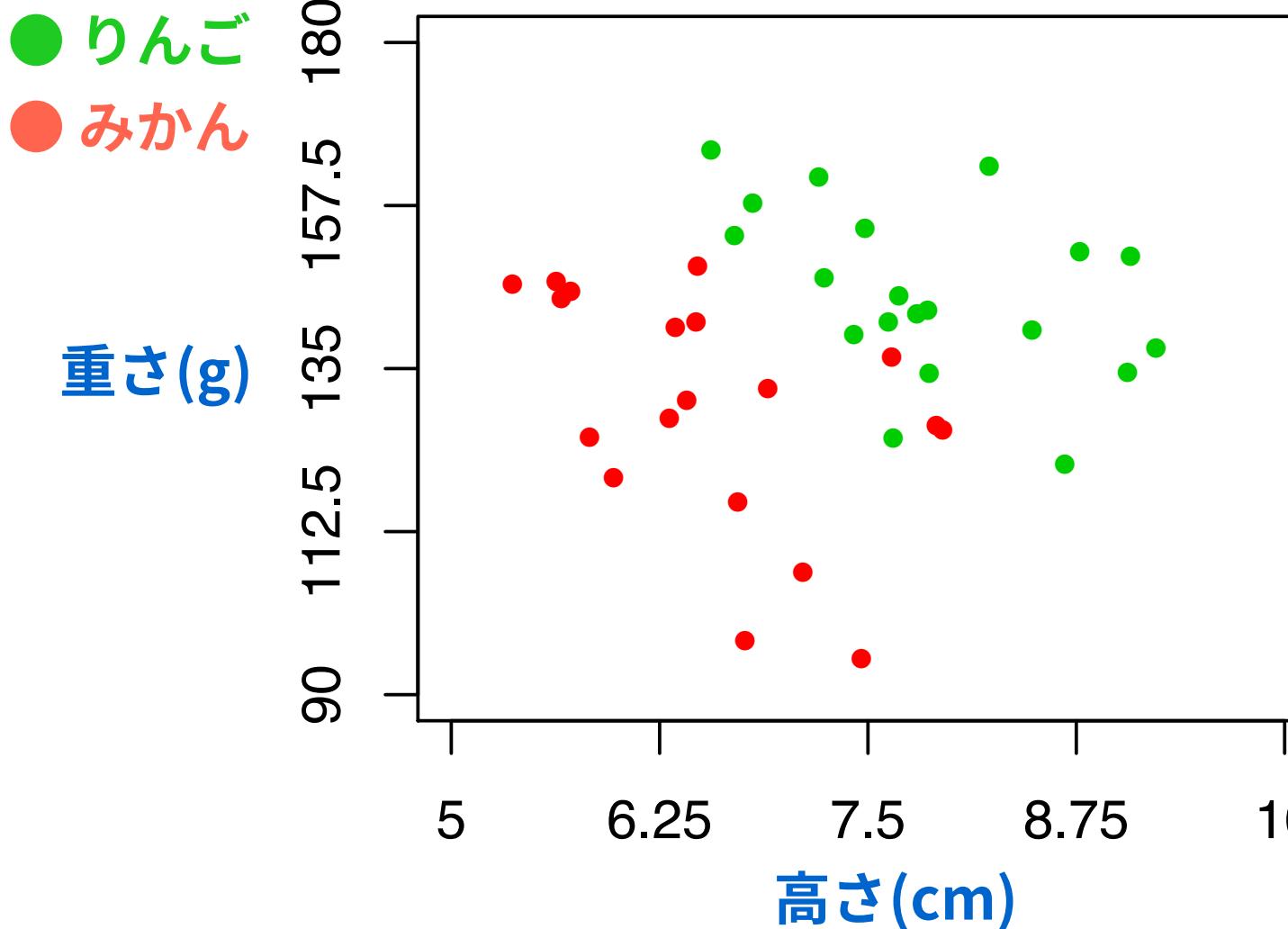
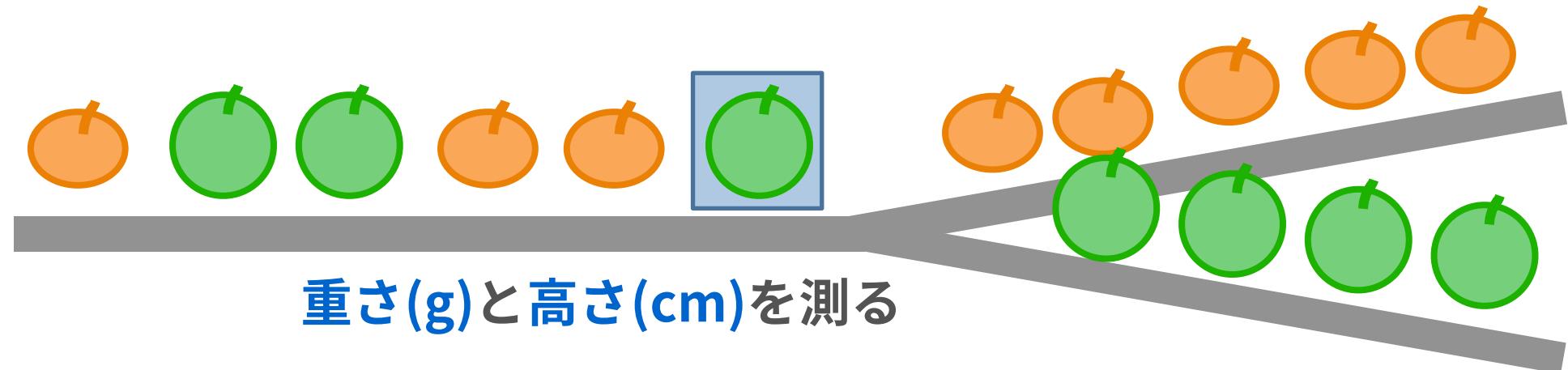
# 機械学習は「データを予測に変える」



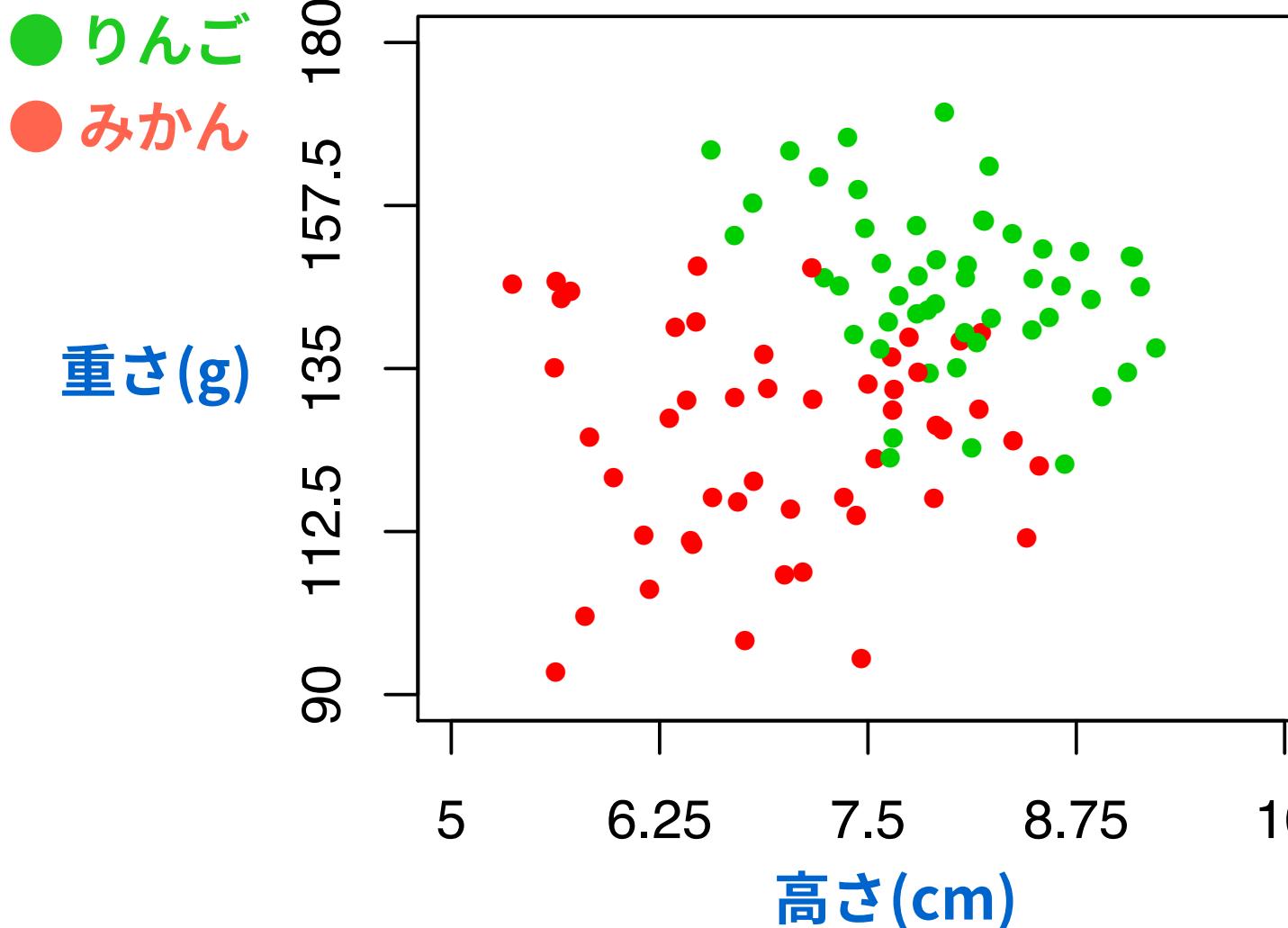
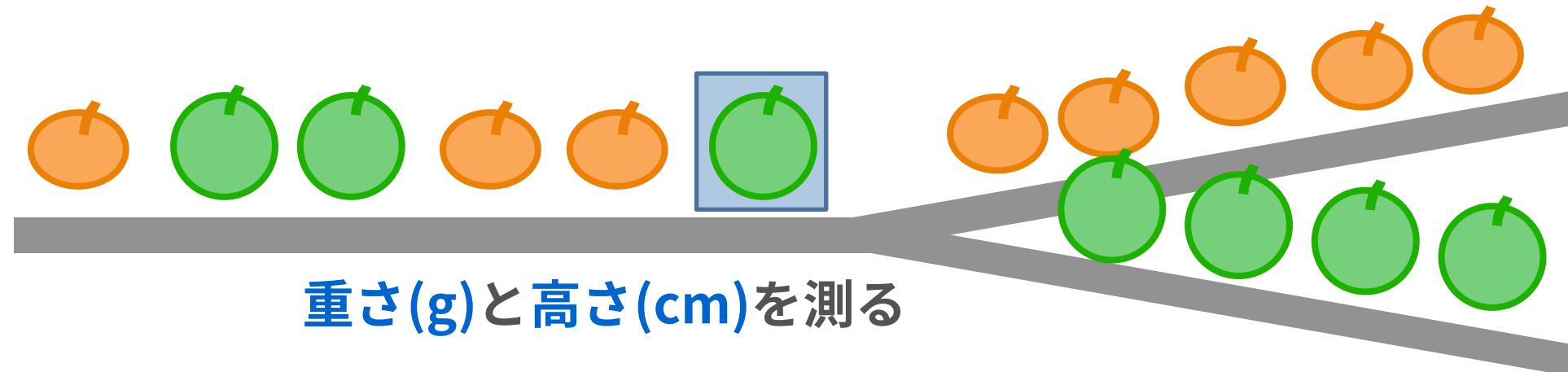
# 機械学習は「データを予測に変える」



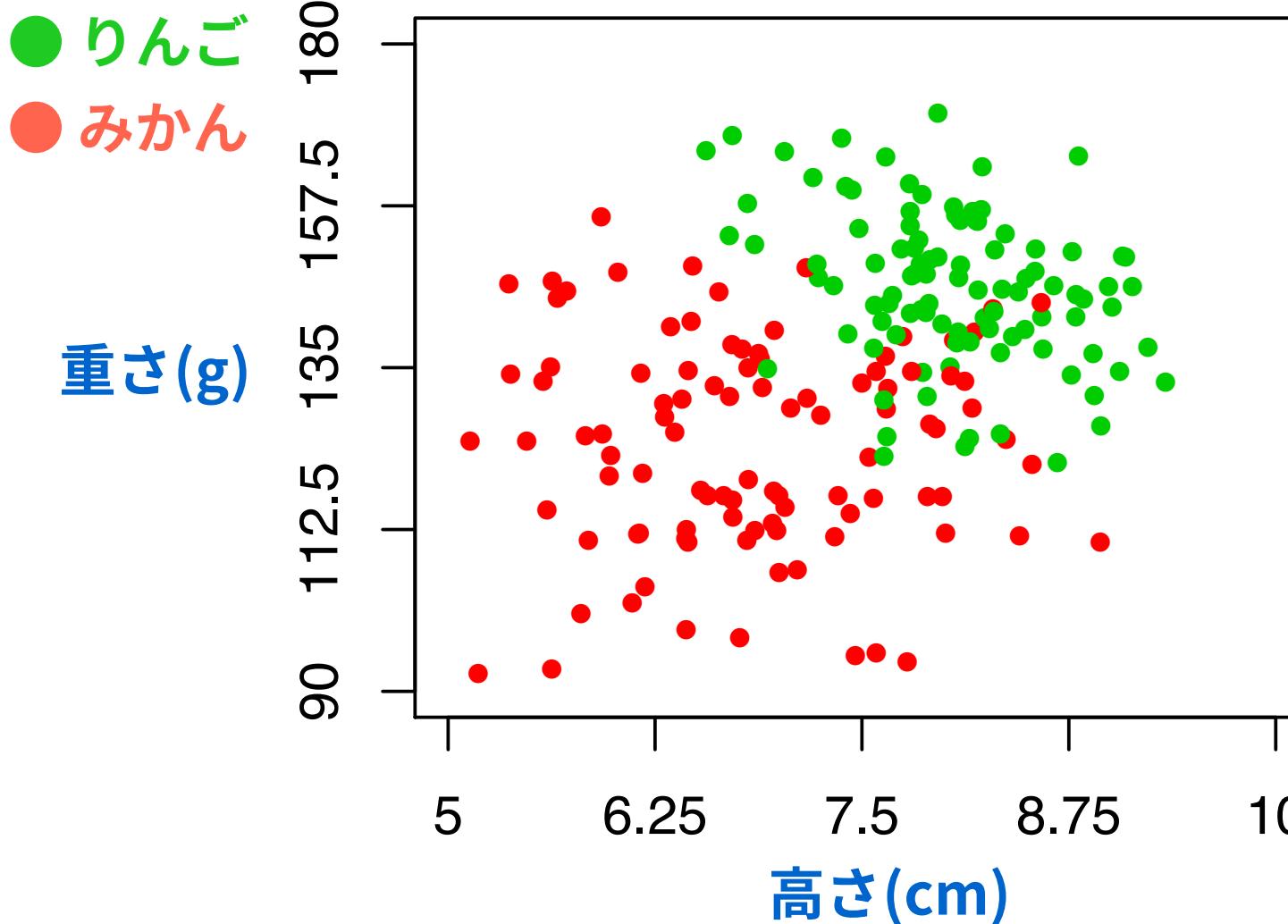
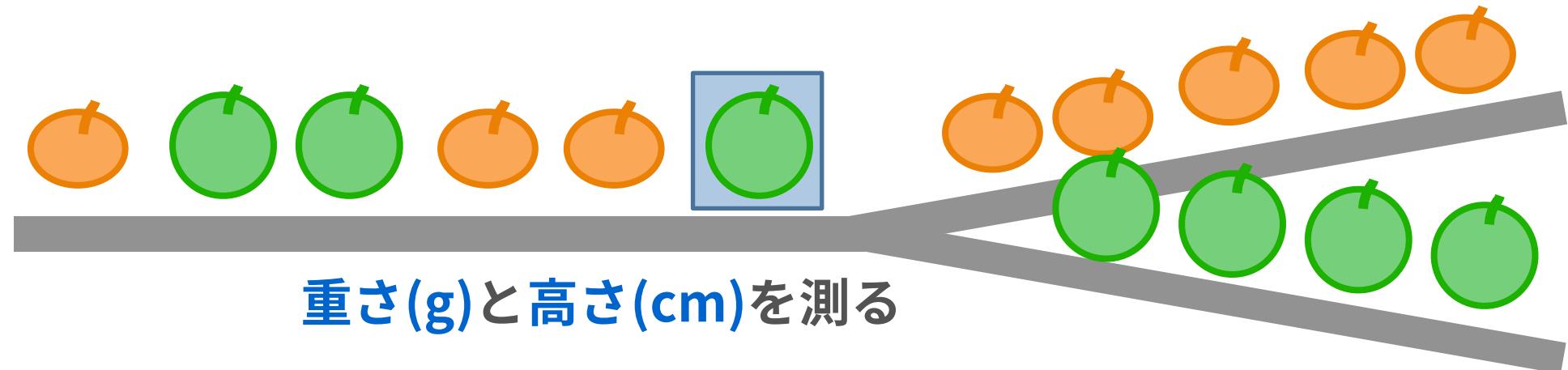
# 機械学習は「データを予測に変える」



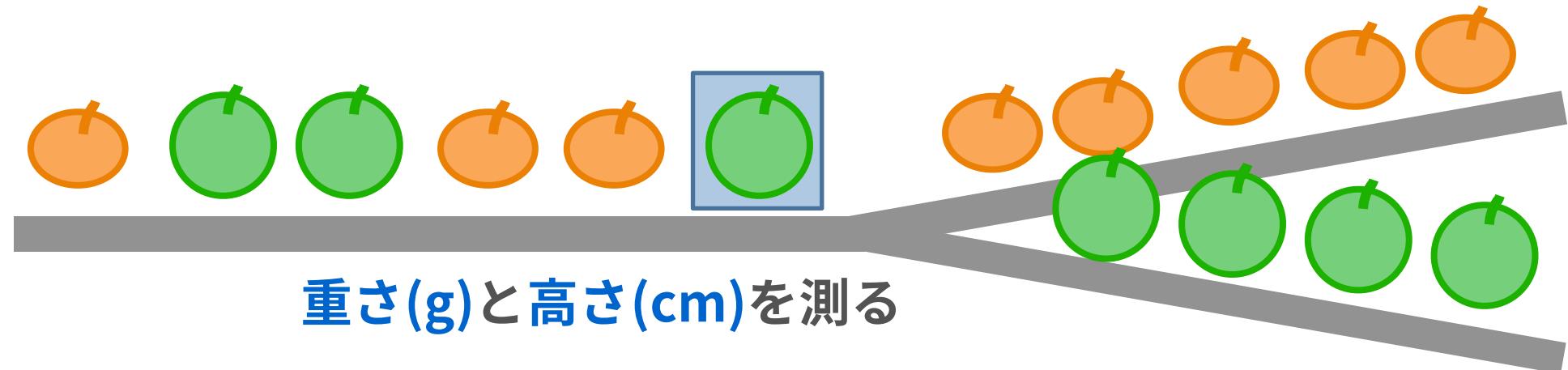
# 機械学習は「データを予測に変える」



# 機械学習は「データを予測に変える」

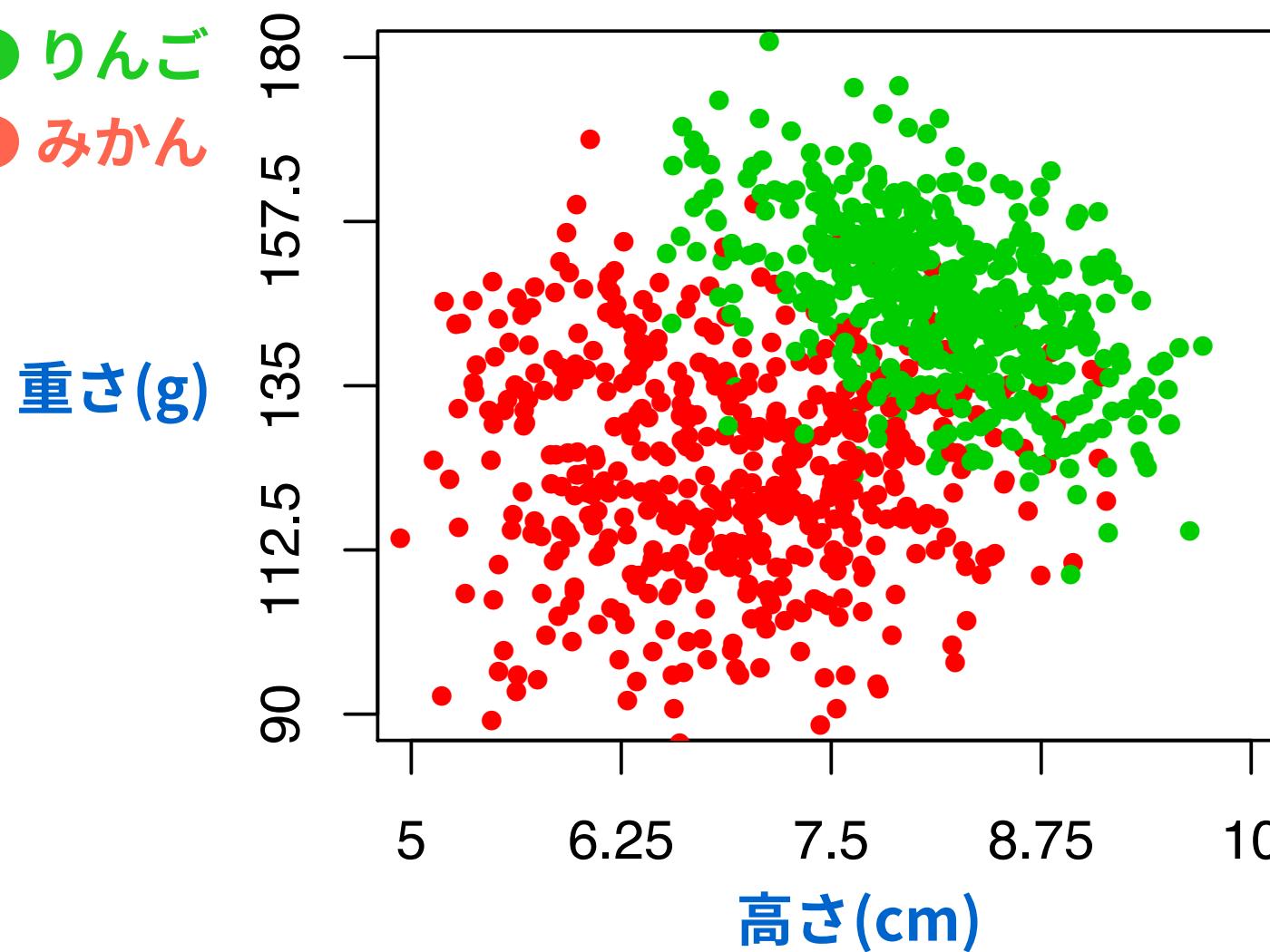


# 機械学習は「データを予測に変える」

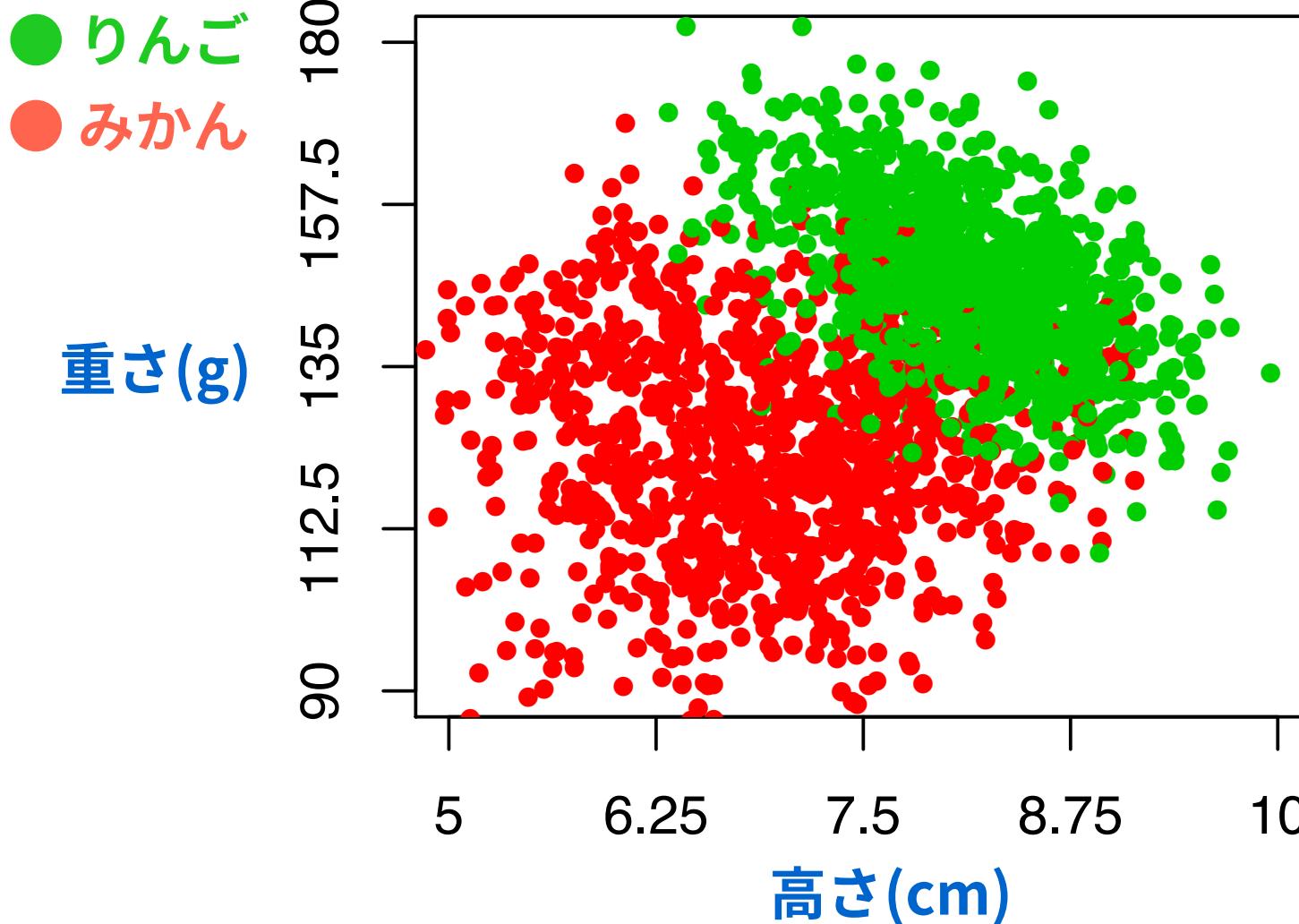
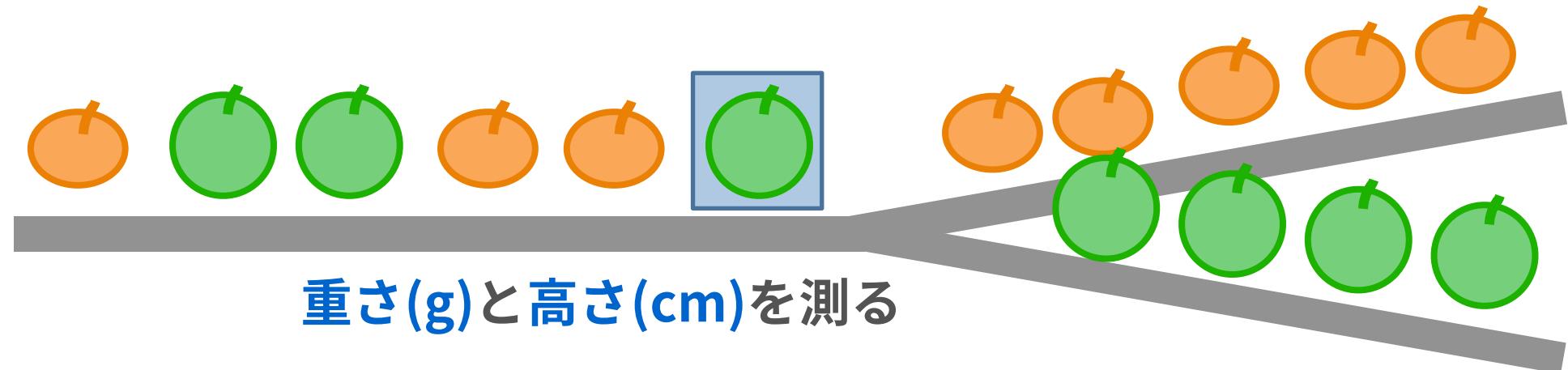


重さ(g)と高さ(cm)を測る

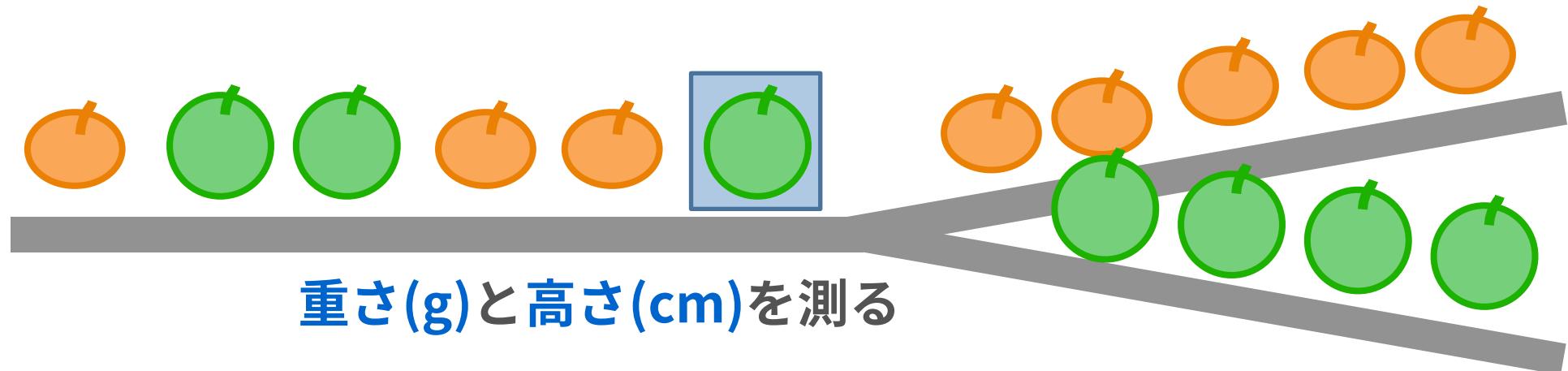
りんご  
みかん



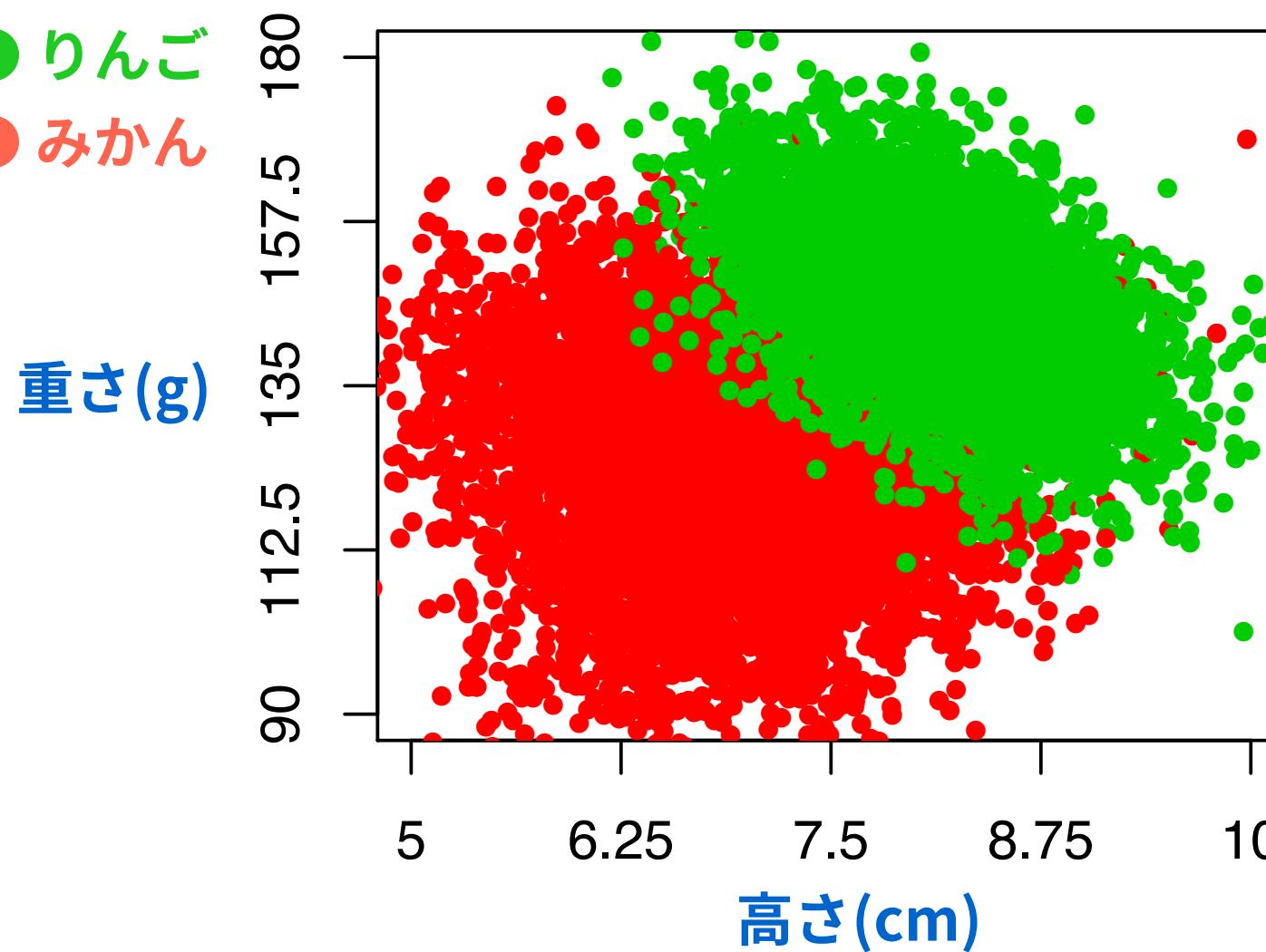
# 機械学習は「データを予測に変える」



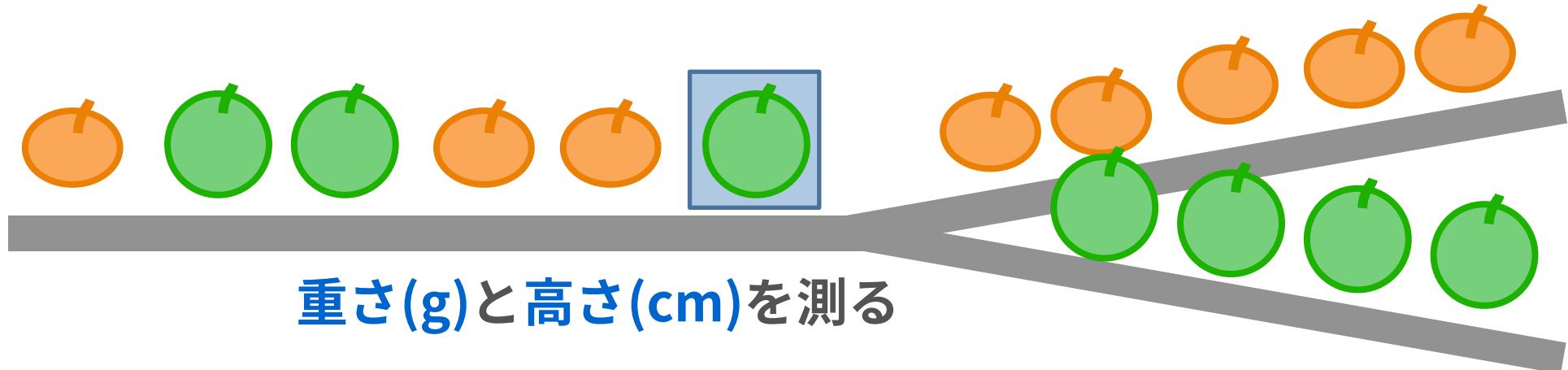
# 機械学習は「データを予測に変える」



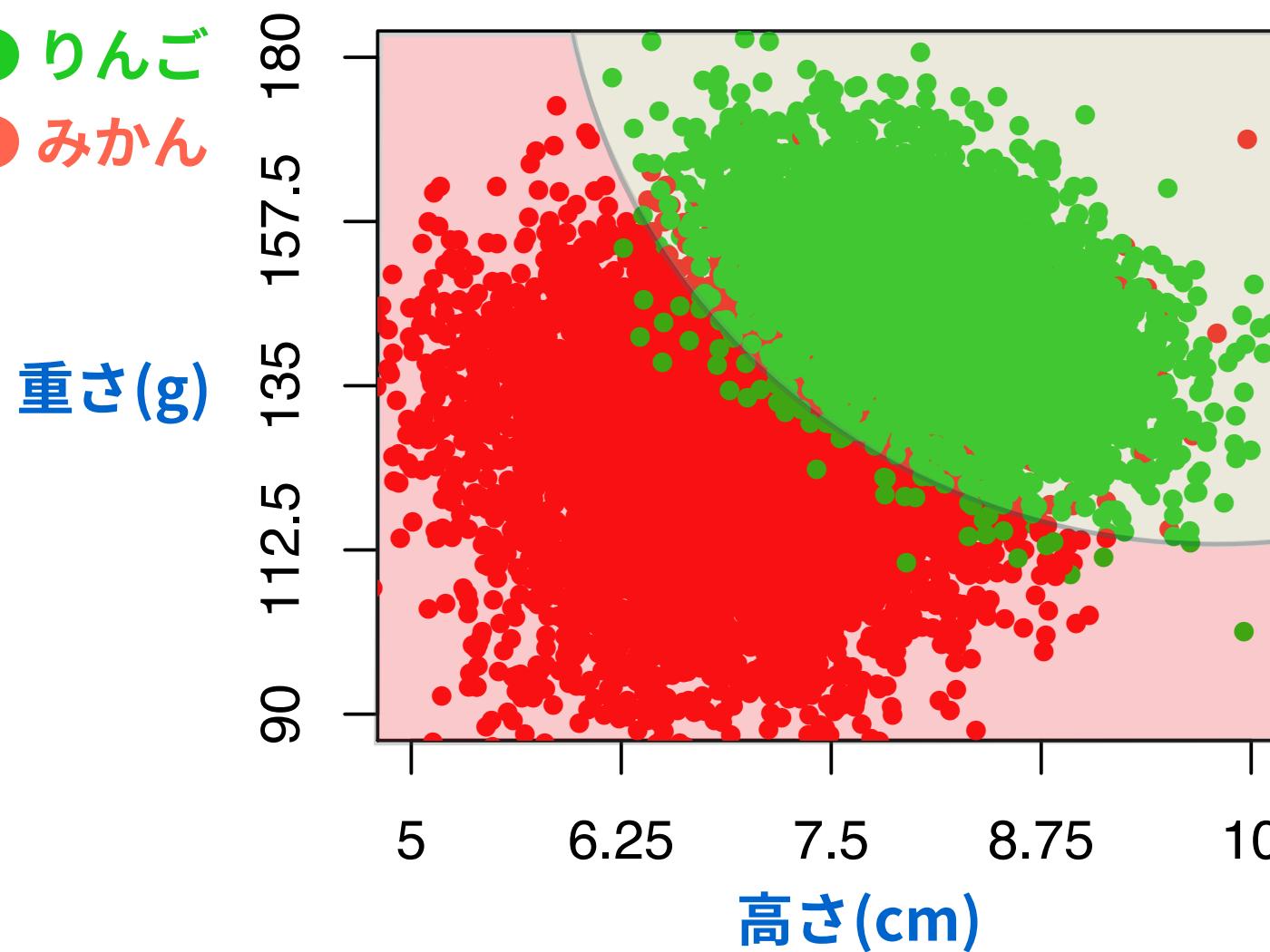
りんご  
みかん



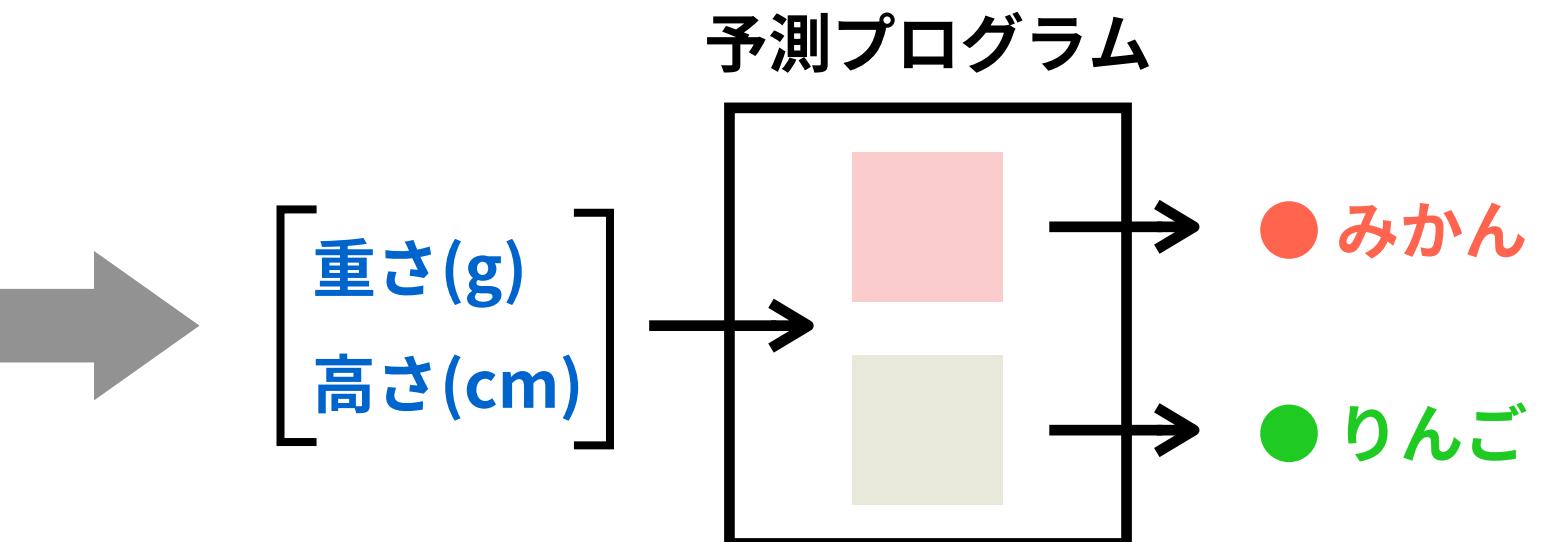
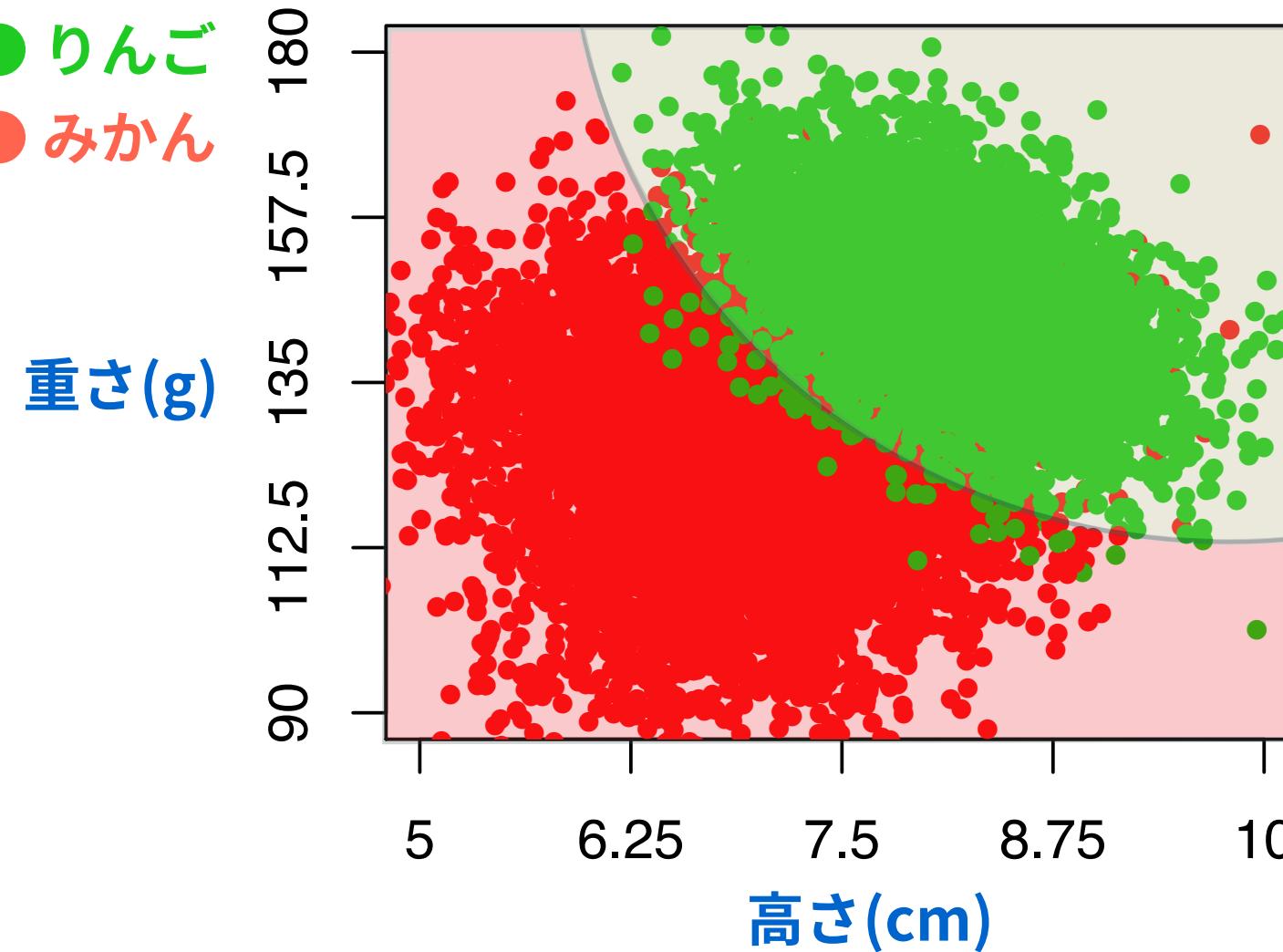
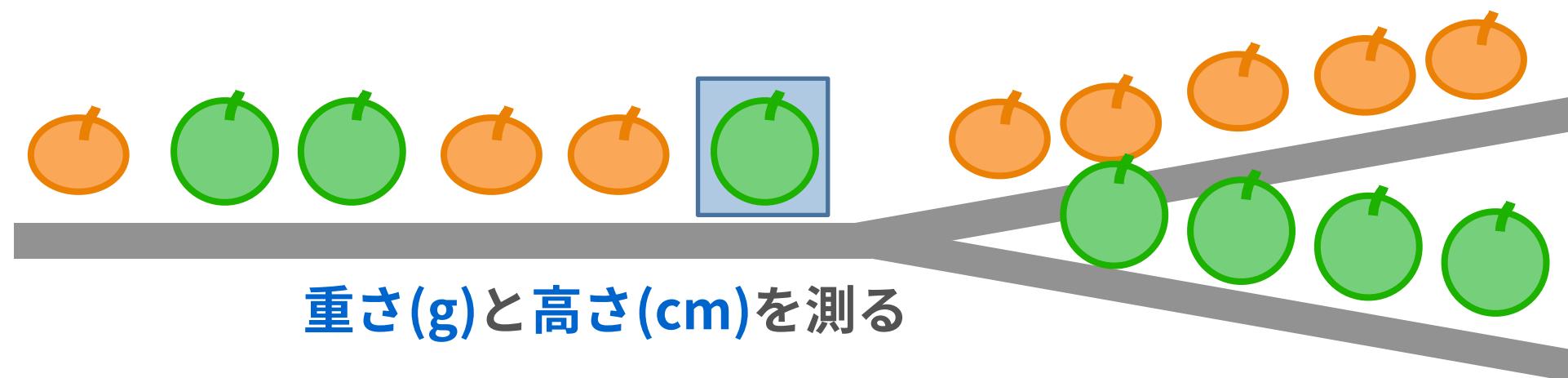
# 機械学習は「データを予測に変える」



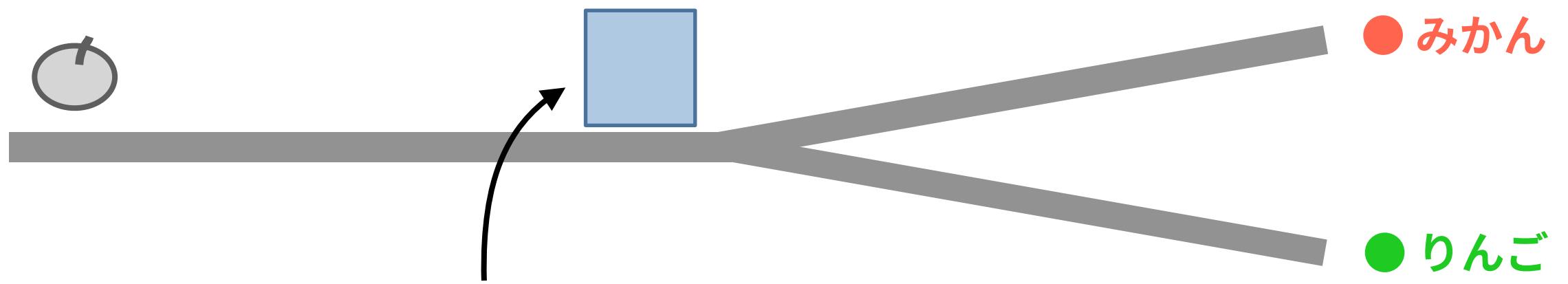
りんご  
みかん



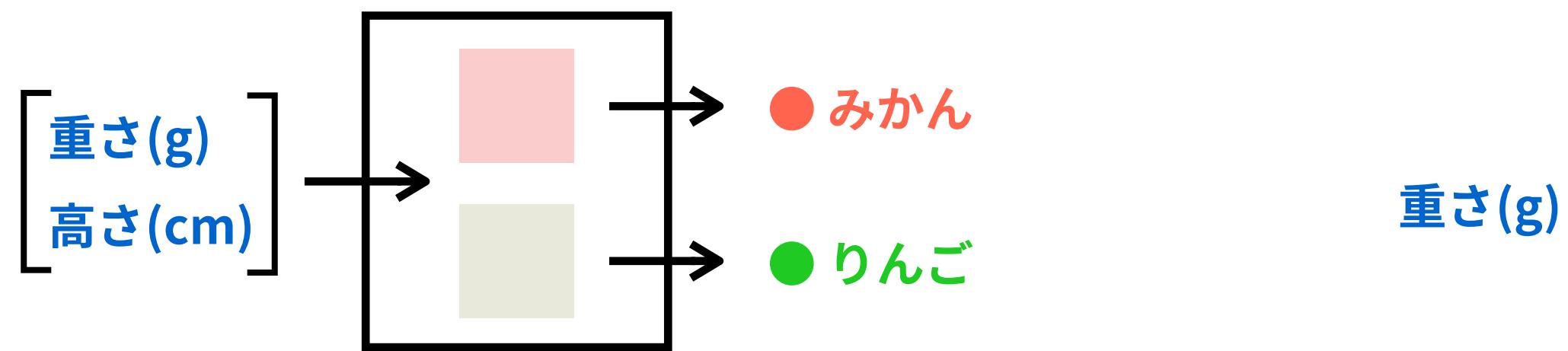
# 機械学習は「データを予測に変える」



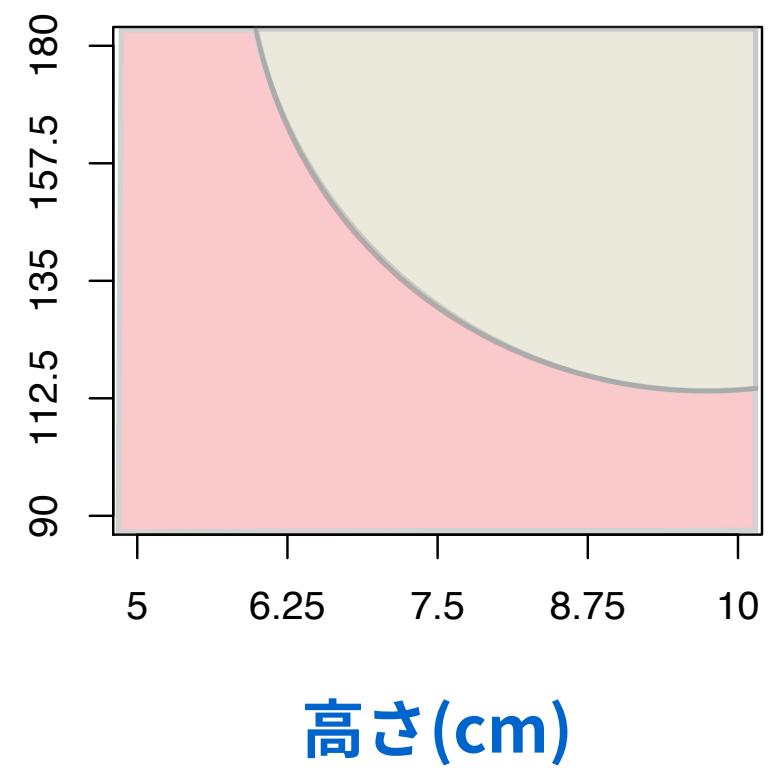
# 機械学習は「データを予測に変える」



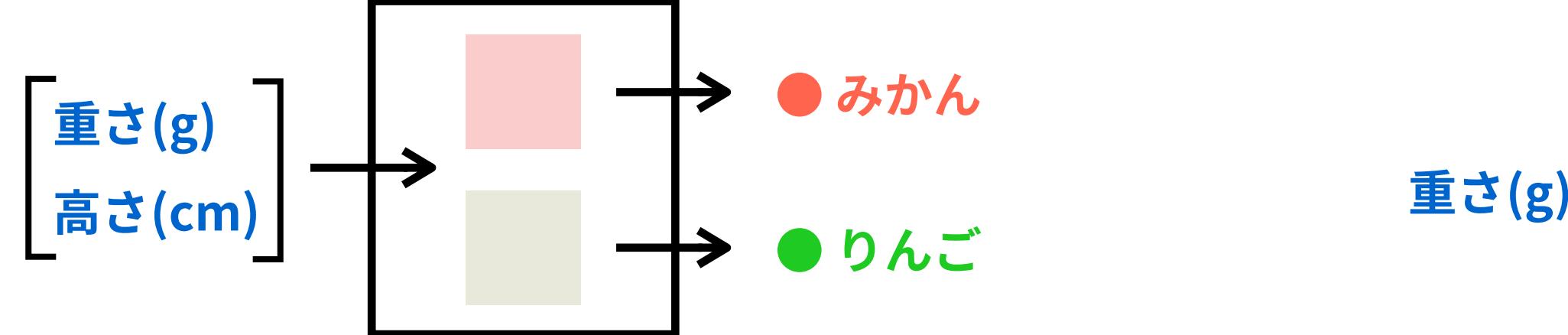
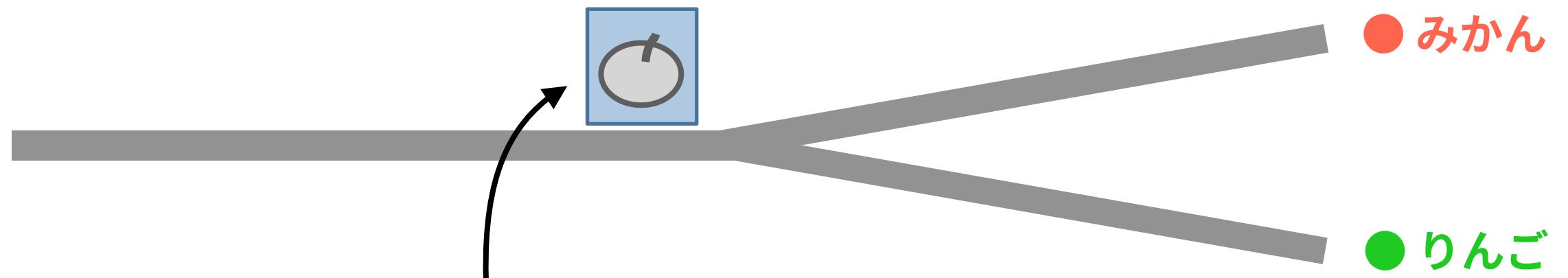
見本データから作っておいた予測プログラム



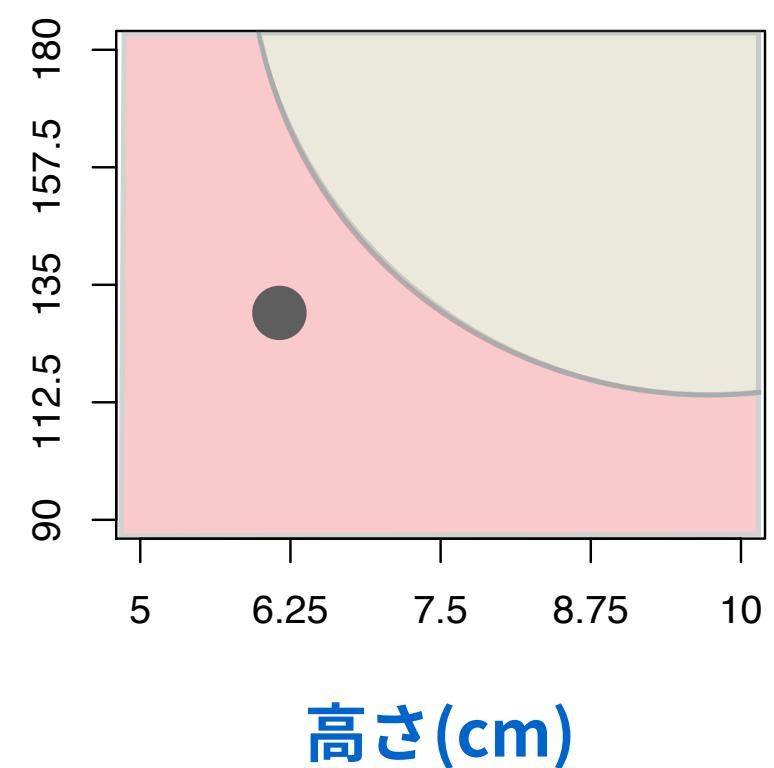
予測プログラムを作ったときに見せた見本例ではない例に対して  
「みかん」 or 「りんご」 を予測することができる！



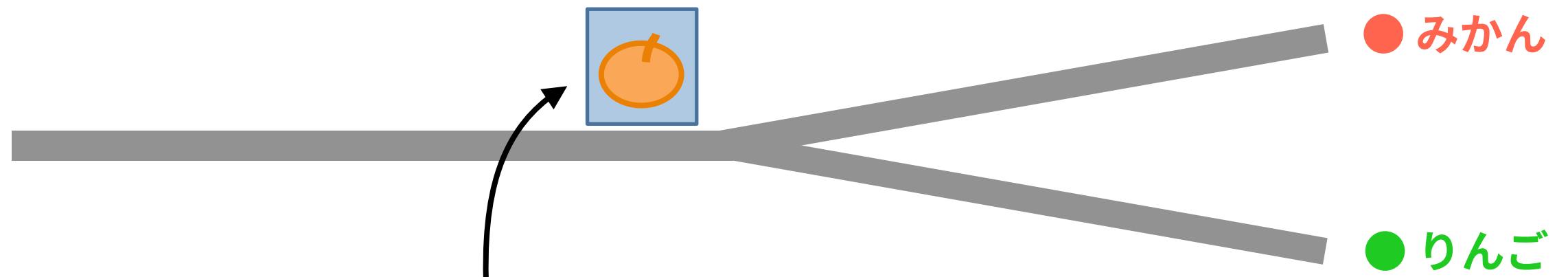
# 機械学習は「データを予測に変える」



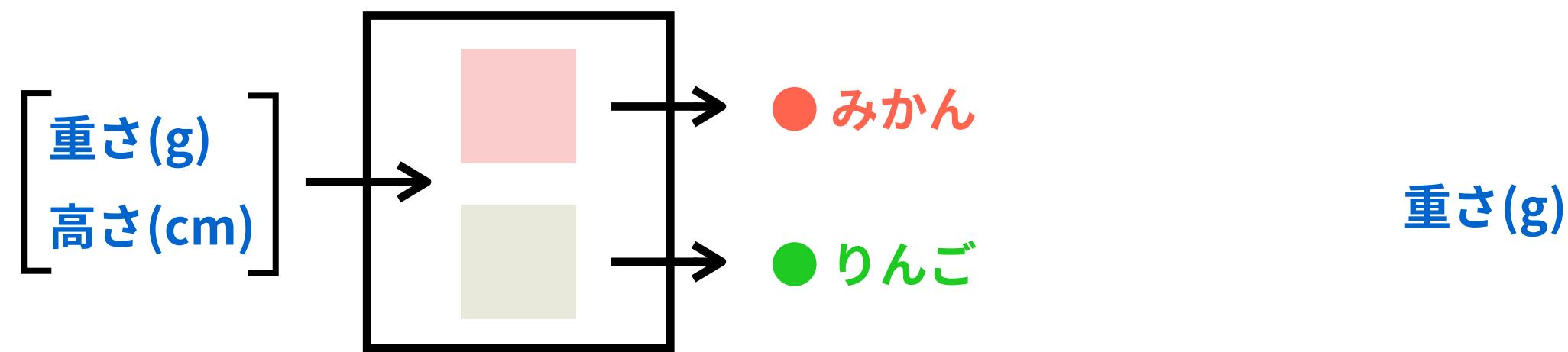
予測プログラムを作ったときに見せた見本例ではない例に対して  
「みかん」 or 「りんご」 を予測することができる！



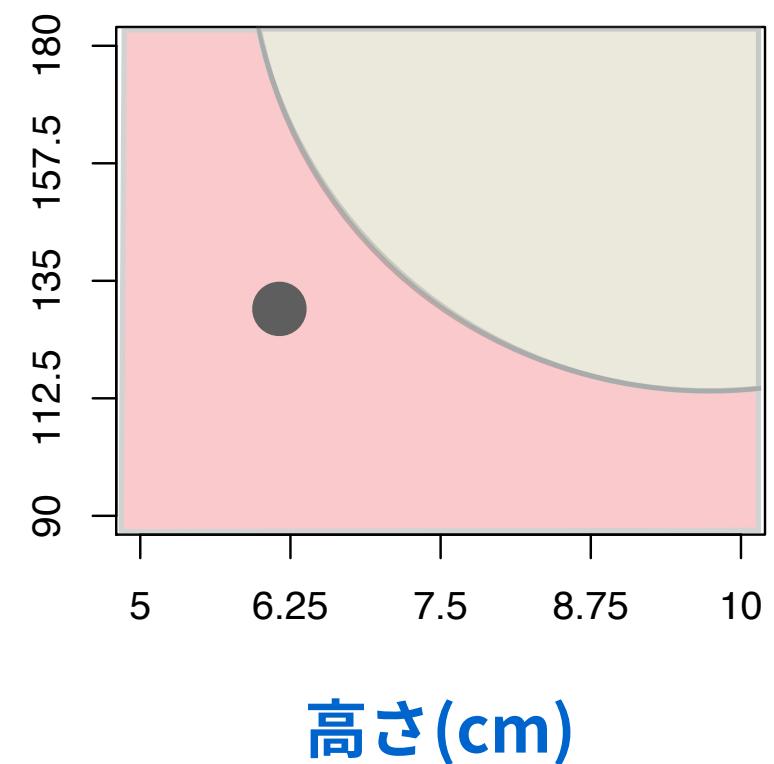
# 機械学習は「データを予測に変える」



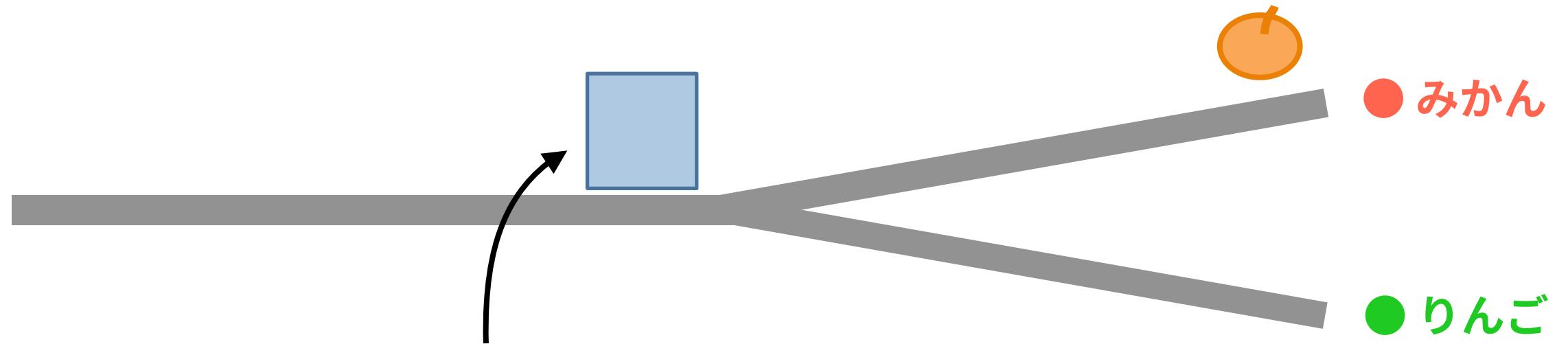
見本データから作っておいた予測プログラム



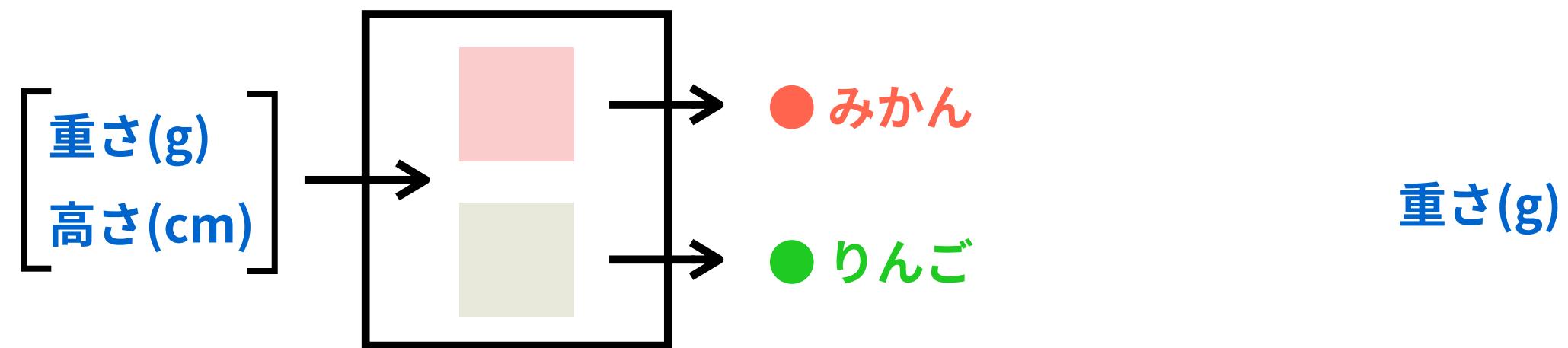
予測プログラムを作ったときに見せた見本例ではない例に対して  
「みかん」 or 「りんご」 を予測することができる！



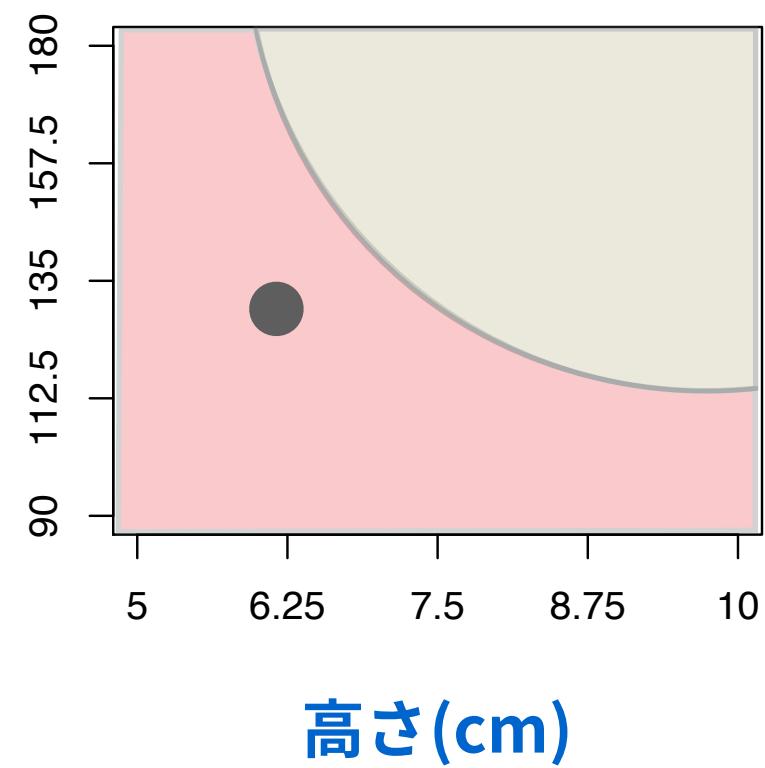
# 機械学習は「データを予測に変える」



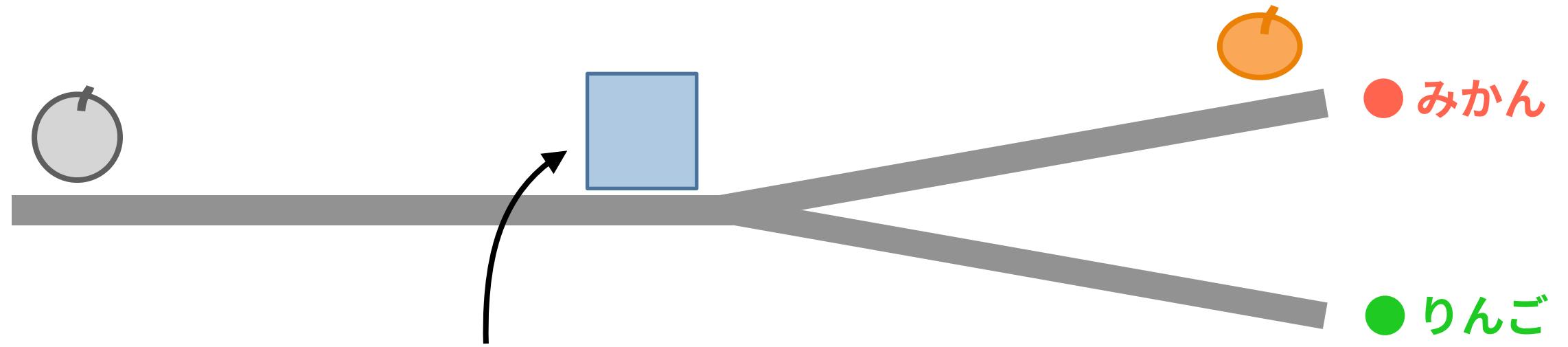
見本データから作っておいた予測プログラム



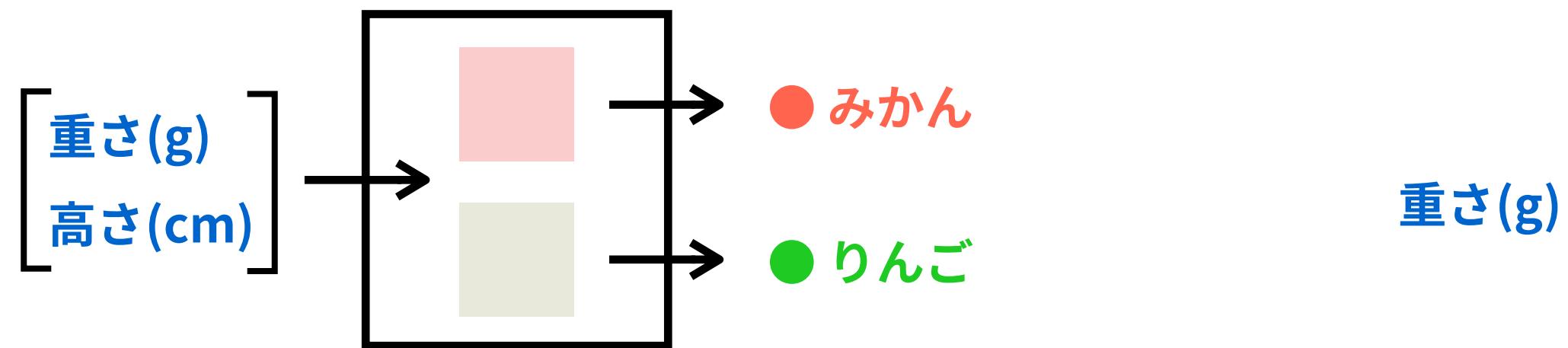
予測プログラムを作ったときに見せた見本例ではない例に対して  
「みかん」 or 「りんご」 を予測することができる！



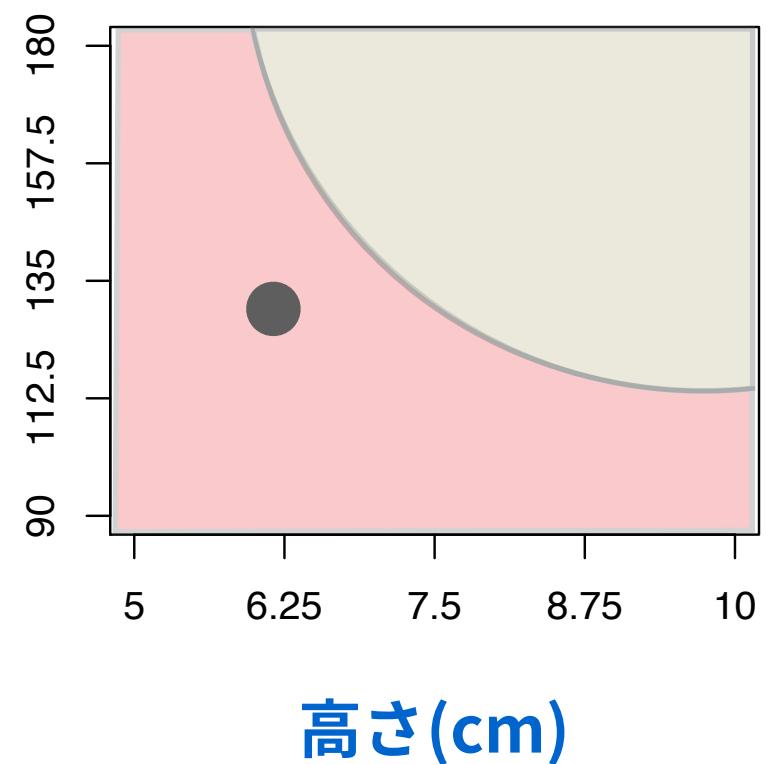
# 機械学習は「データを予測に変える」



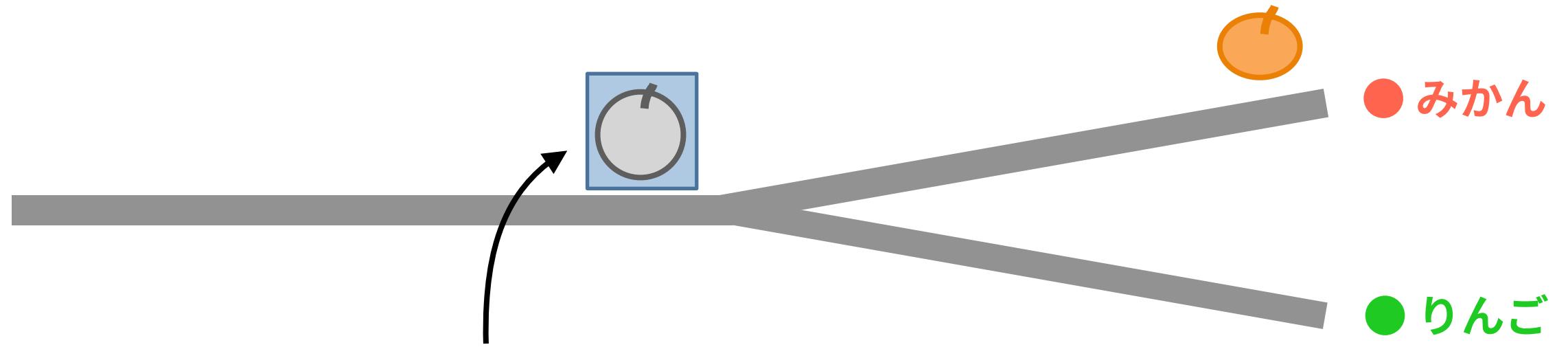
見本データから作っておいた予測プログラム



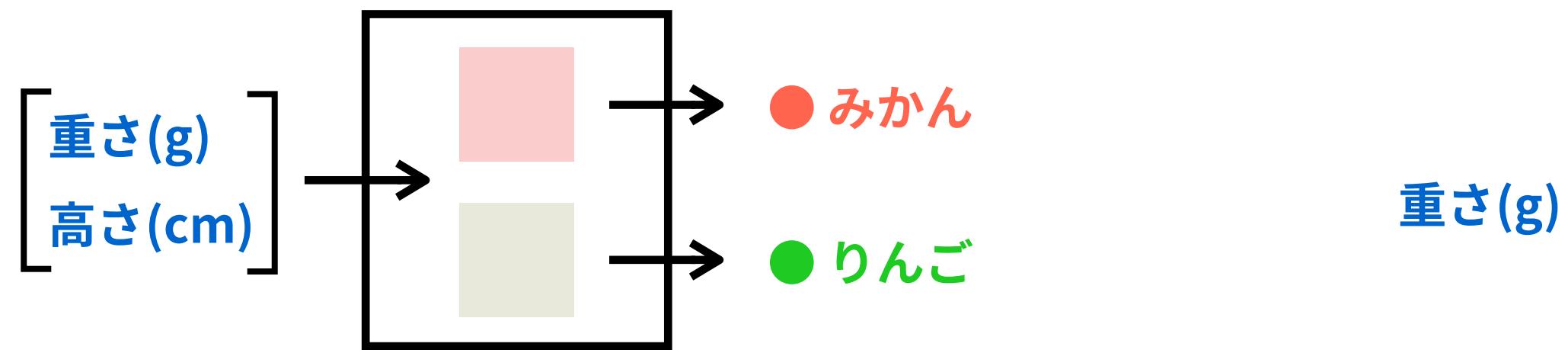
予測プログラムを作ったときに見せた見本例ではない例に対して  
「みかん」 or 「りんご」を予測することができる！



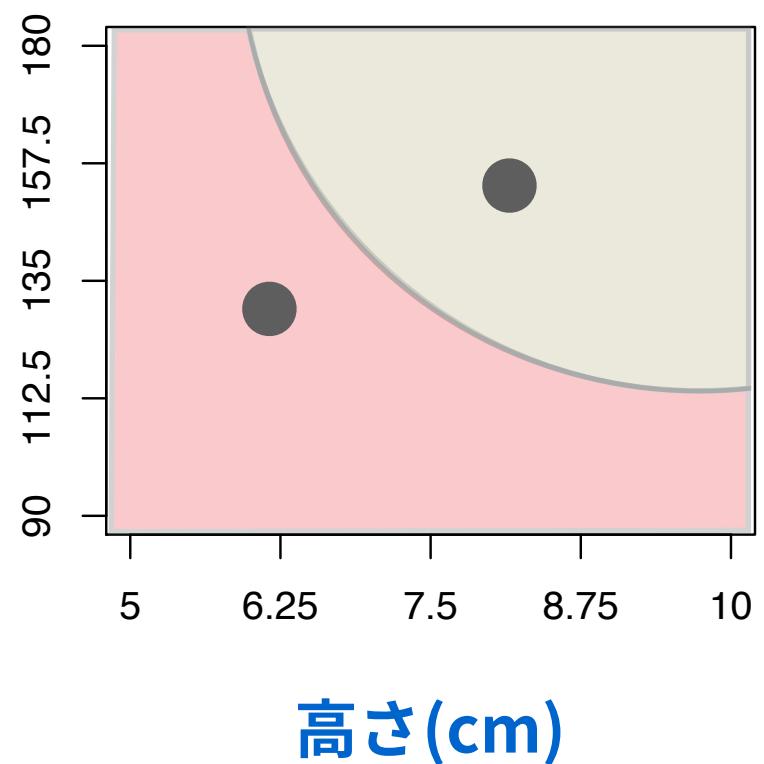
# 機械学習は「データを予測に変える」



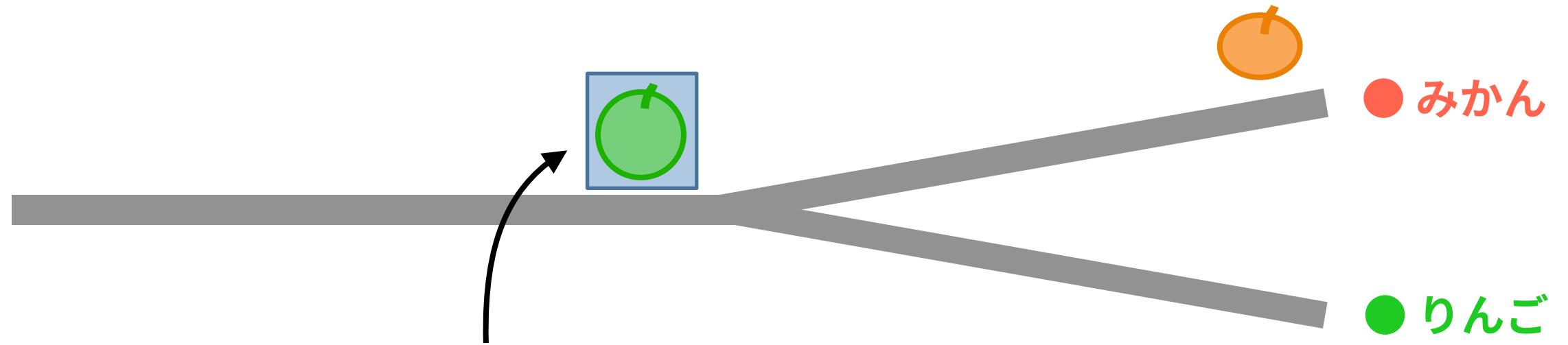
見本データから作っておいた予測プログラム



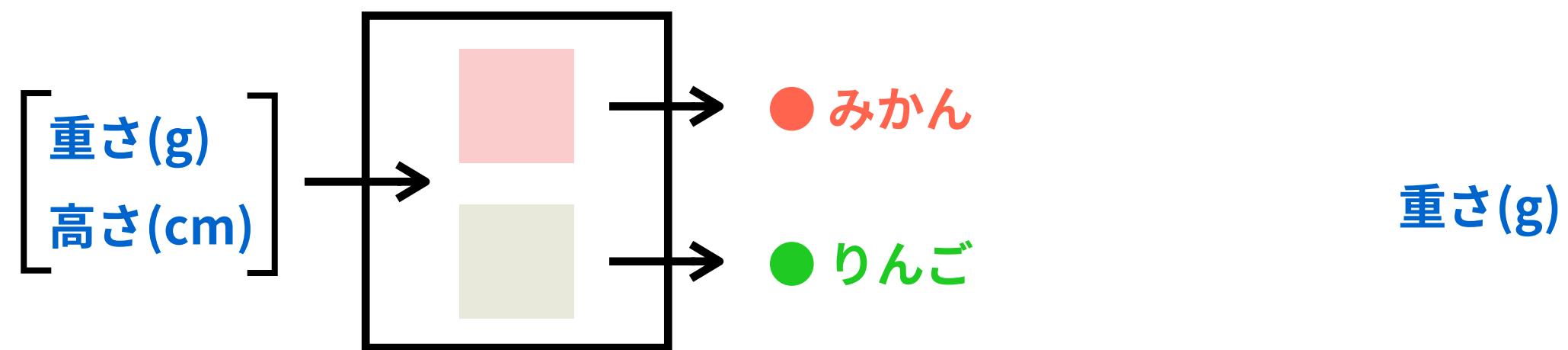
予測プログラムを作ったときに見せた見本例ではない例に対して  
**「みかん」 or 「りんご」** を予測することができる！



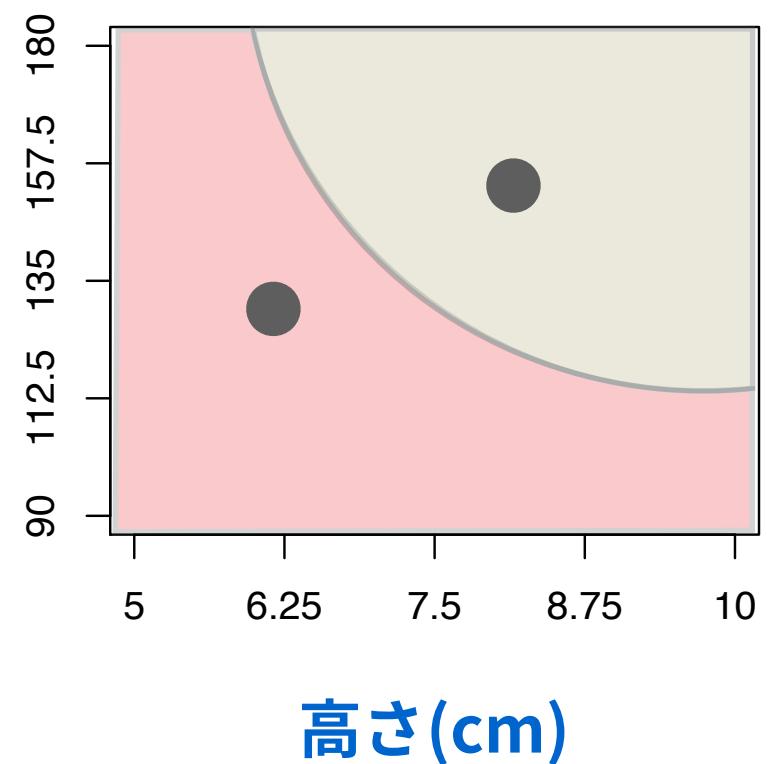
# 機械学習は「データを予測に変える」



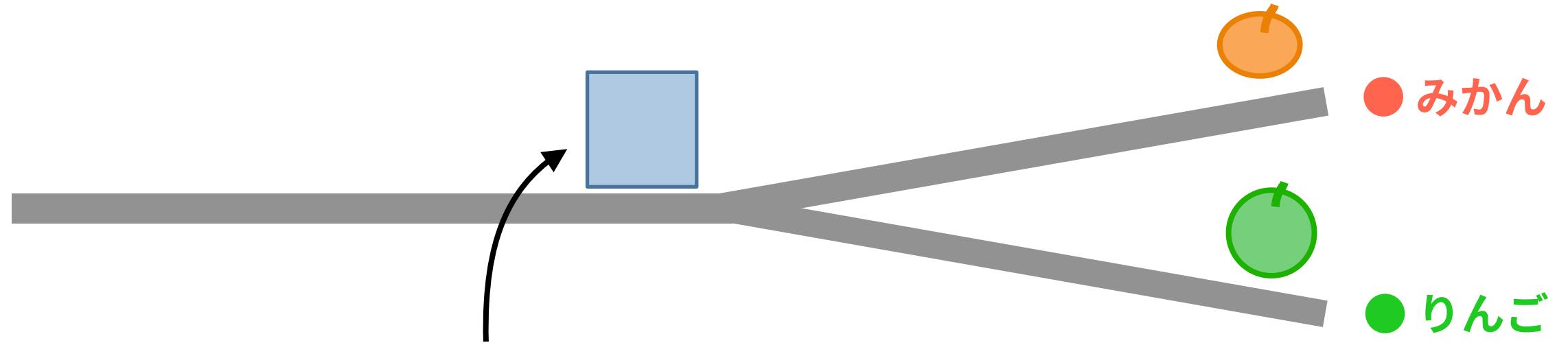
見本データから作っておいた予測プログラム



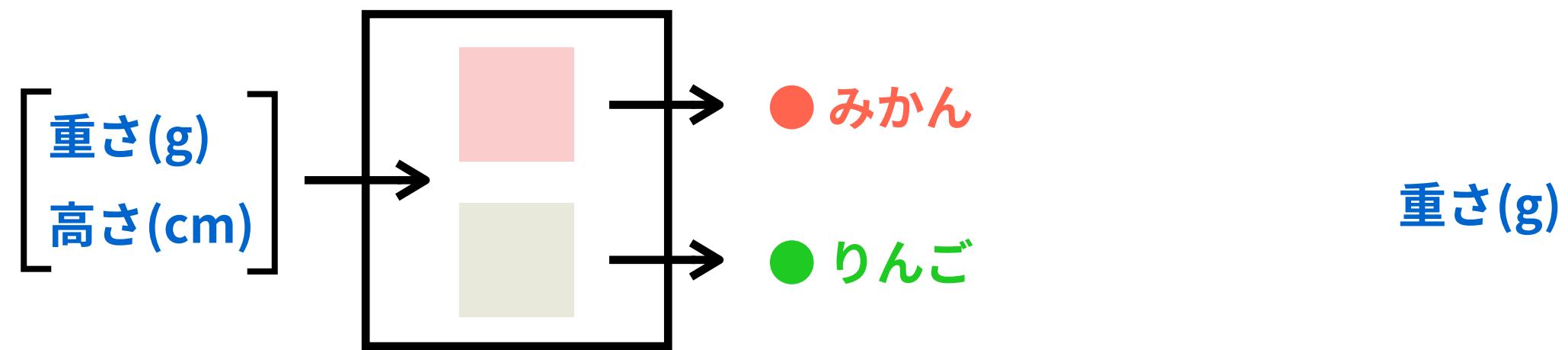
予測プログラムを作ったときに見せた見本例ではない例に対して  
「みかん」 or 「りんご」 を予測することができる！



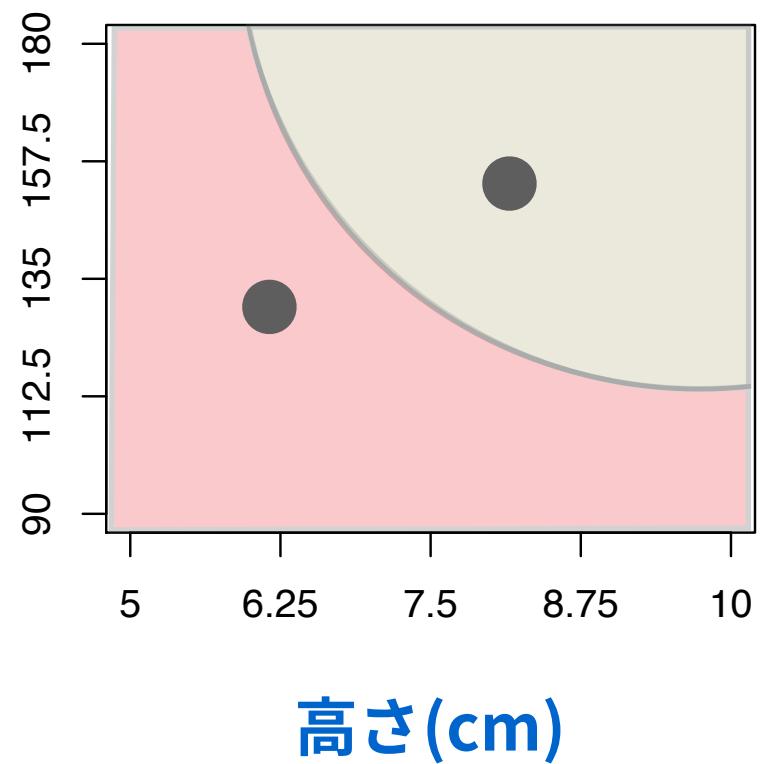
# 機械学習は「データを予測に変える」



見本データから作っておいた予測プログラム



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



# 機械学習は「新しいコンピュータプログラムの作り方」

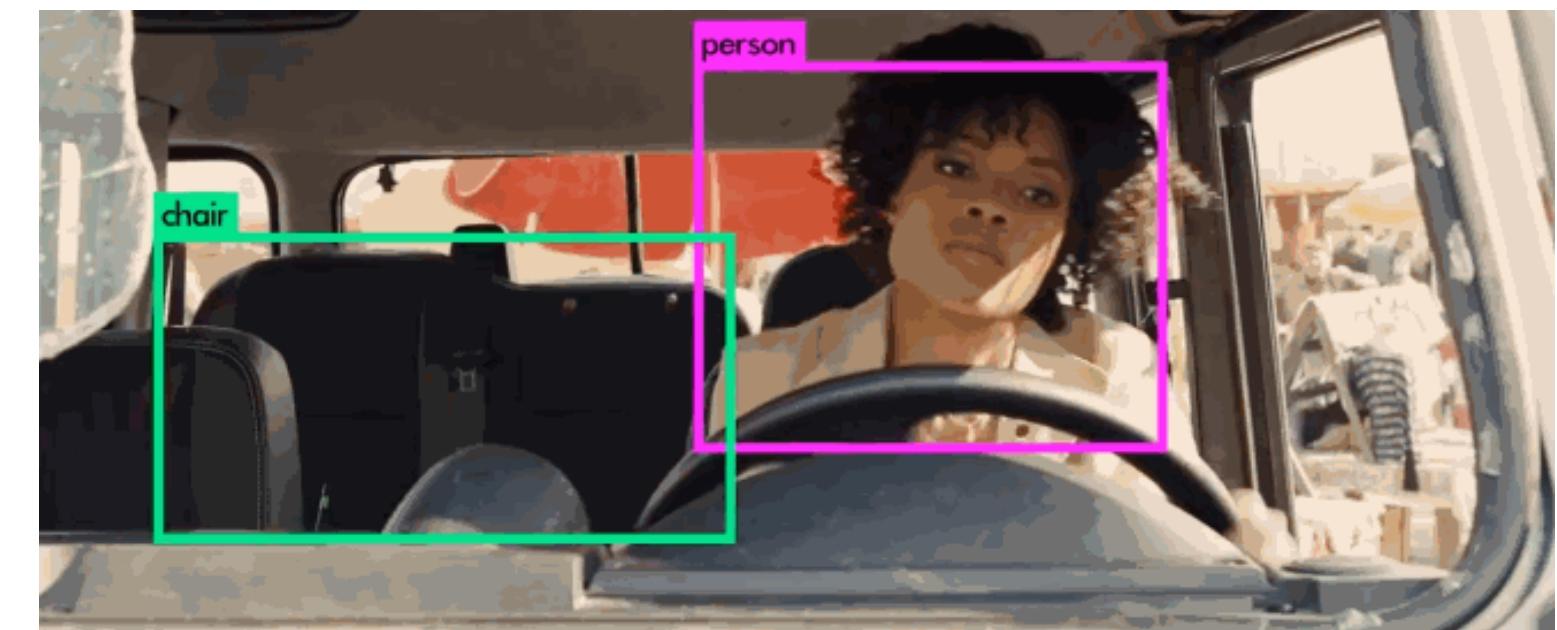


プログラムの入力と出力の関係がよく分からない場合でも、  
たくさんの入出力の見本データによって間接的にそれを再現できるプログラムを作り出す技術



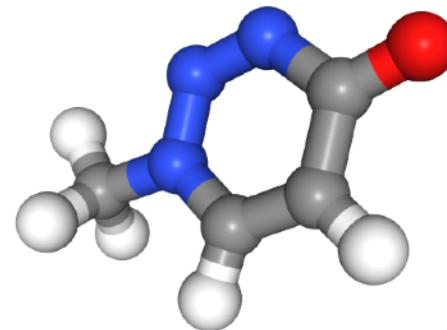
# 機械学習は「新しいコンピュータプログラムの作り方」

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



# 例：機械學習×量子化学

input



gdb\_21014

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

~ 1000 秒

量子化学計算

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

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

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

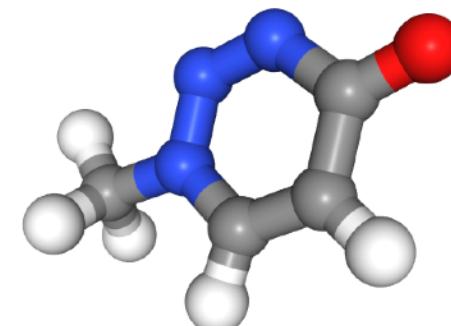
output

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

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

# 例：機械學習×量子化学

input



gdb\_21014

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

100,000 倍高速！

~ 0.01 秒

機械學習



~ 1000 秒

量子化学計算

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

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

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

output

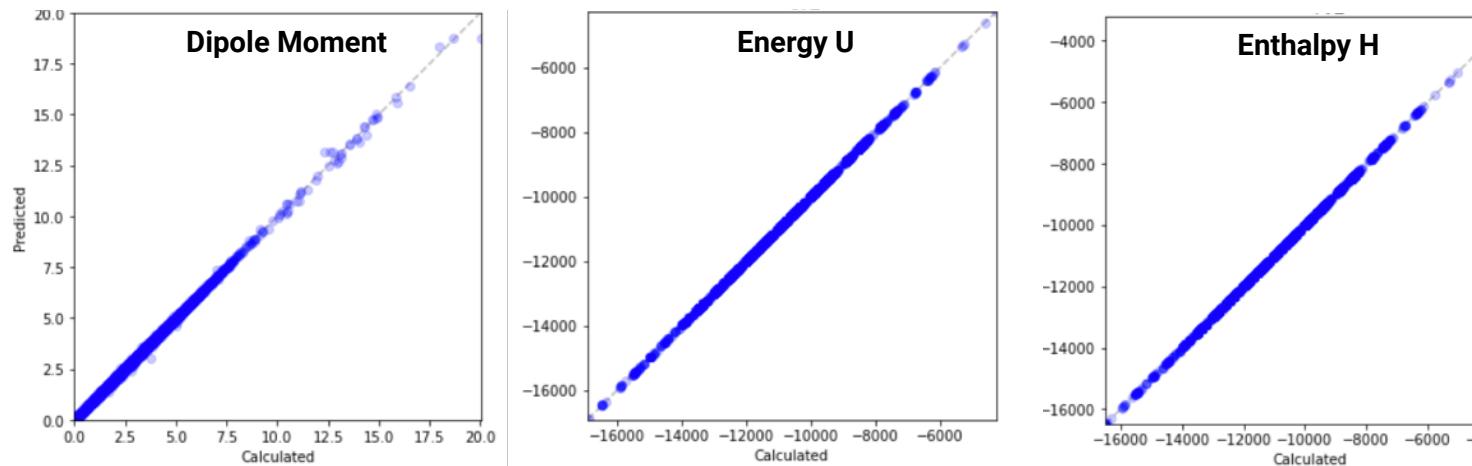
	property	value
0	dipole_moment	7.214000
1	isotropic_polarizability	65.360001
2	homo	-6.280388
3	lumo	-1.649010
4	gap	4.631378
5	electronic_spatial_extent	884.587524
6	zpve	2.610307
7	energy_U0	-10742.250000
8	energy_U	-10742.060547
9	enthalpy_H	-10742.035156
10	free_energy_G	-10743.111328
11	heat_capacity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

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

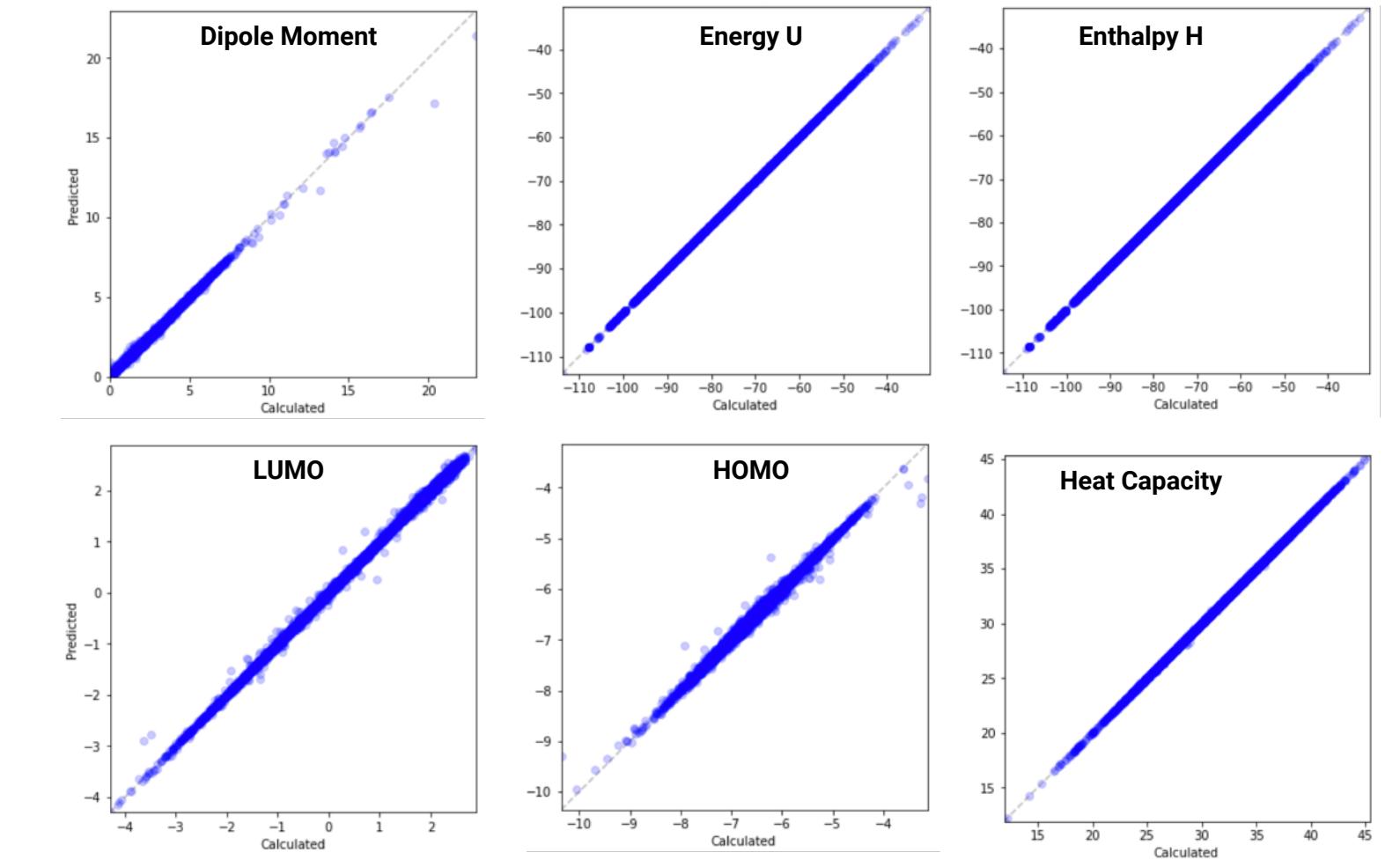
# 例：機械学習×量子化学

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

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



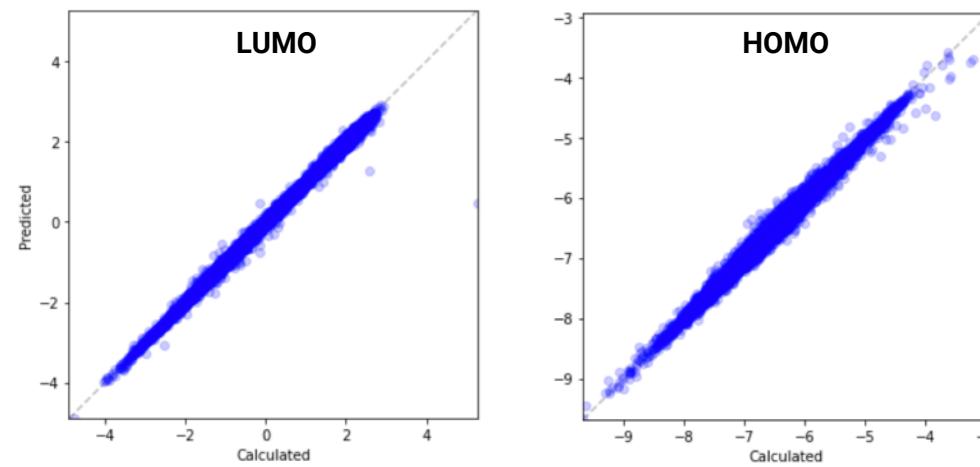
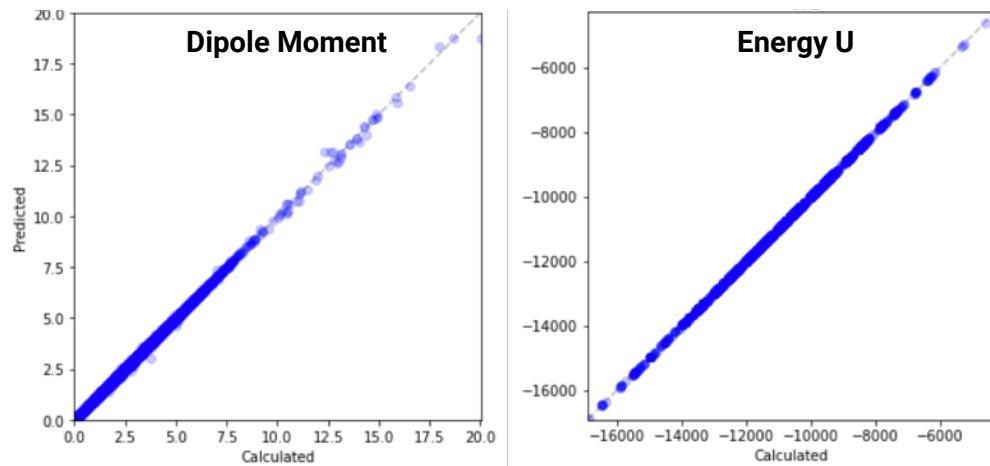
真値(x軸) vs 予測値(y軸) by DimeNet (Klicpera et al, 2020)



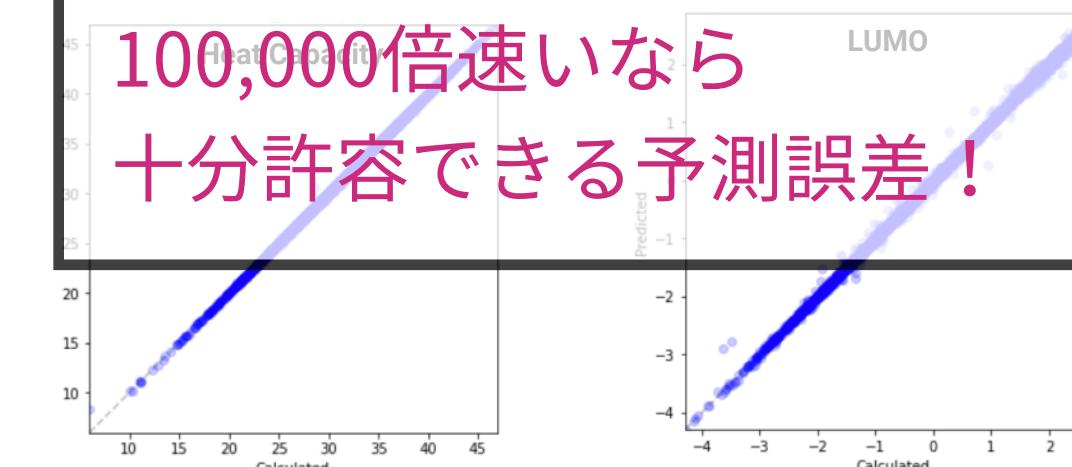
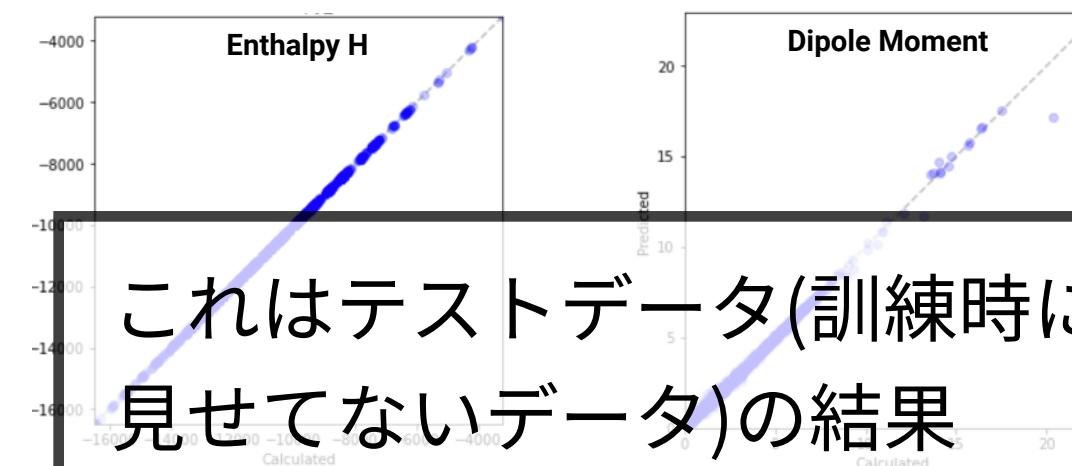
# 例：機械学習×量子化学

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

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



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

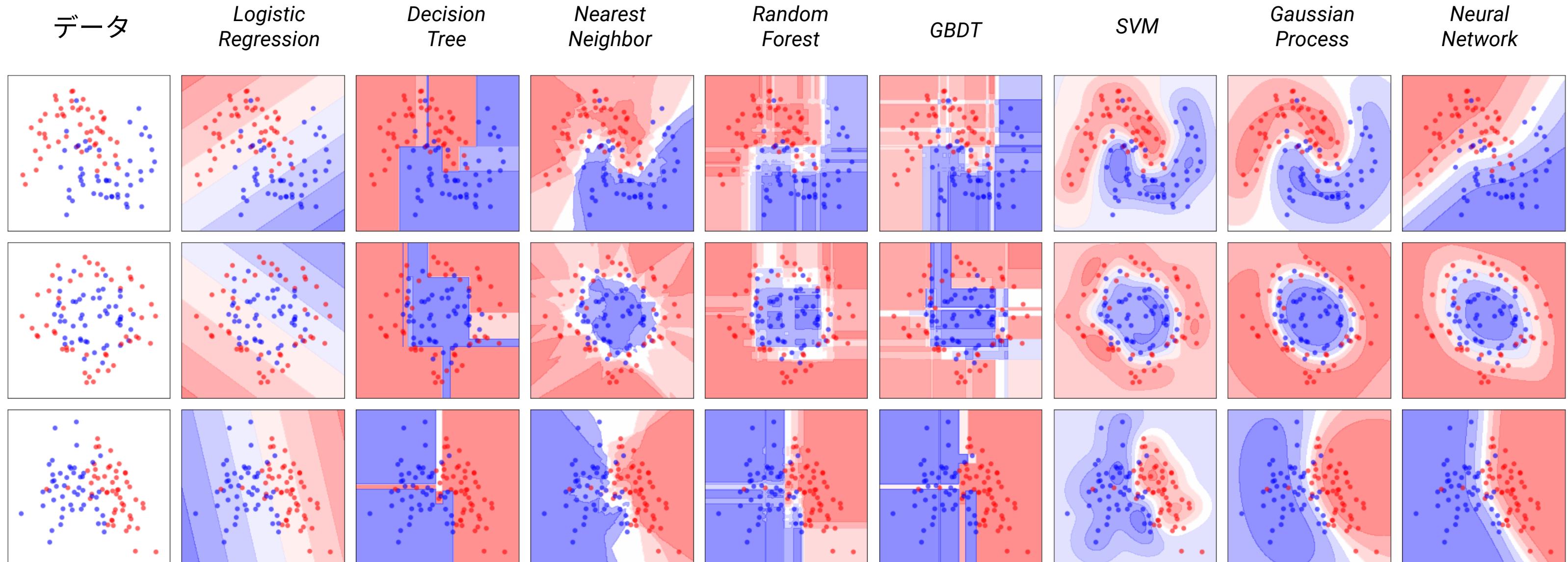


これはテストデータ(訓練時に  
見せてないデータ)の結果

100,000倍速いなら  
十分許容できる予測誤差！

# 機械学習は「新しいコンピュータプログラムの作り方」

機械学習のアルゴリズムの違いは**境界線の引き方の方針のちがい**だけ



# 機械学習は「新しいコンピュータプログラムの作り方」

内部原理は「曲面モデル」を点にフィッティングしているだけ

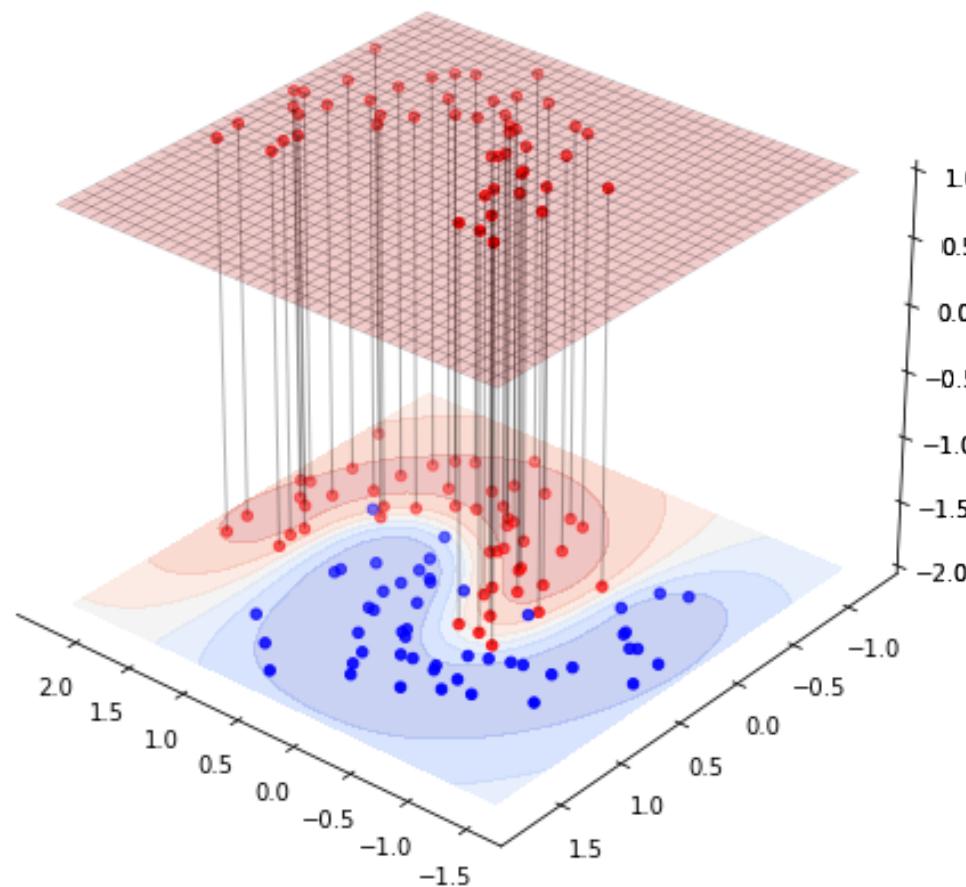
→ 境界線の引き方の方針



# 機械学習は「新しいコンピュータプログラムの作り方」

内部原理は「曲面モデル」を点にフィッティングしているだけ

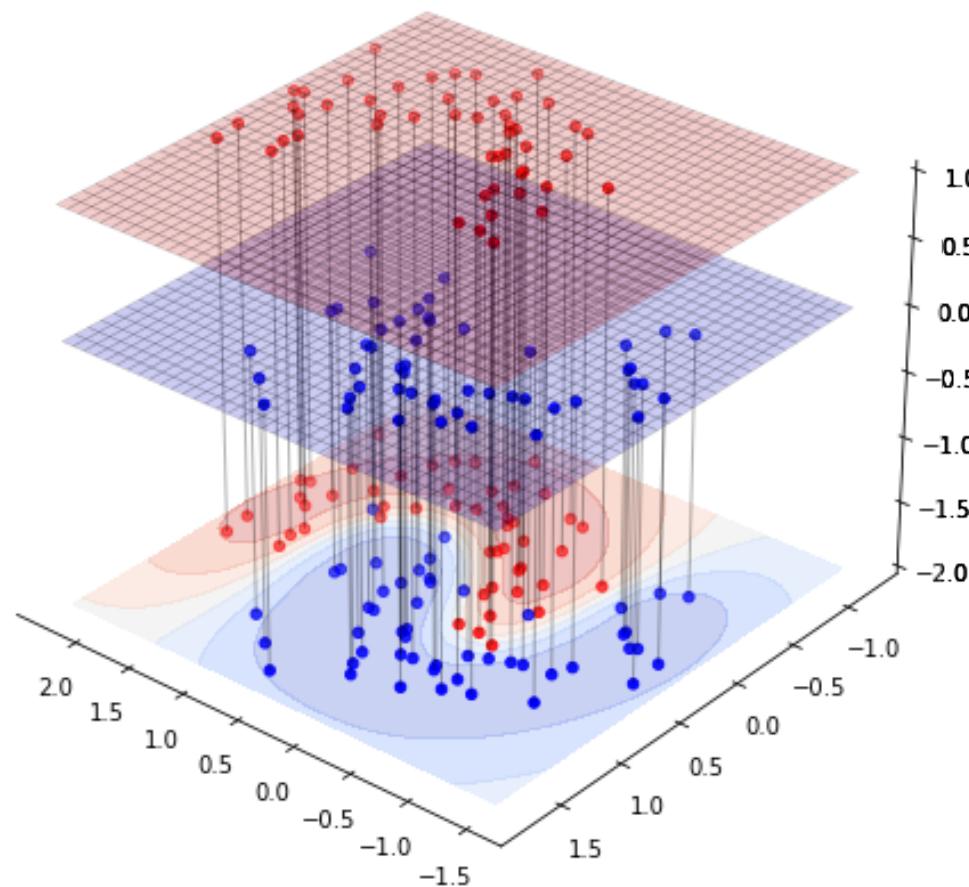
→ 境界線の引き方の方針



# 機械学習は「新しいコンピュータプログラムの作り方」

内部原理は「曲面モデル」を点にフィッティングしているだけ

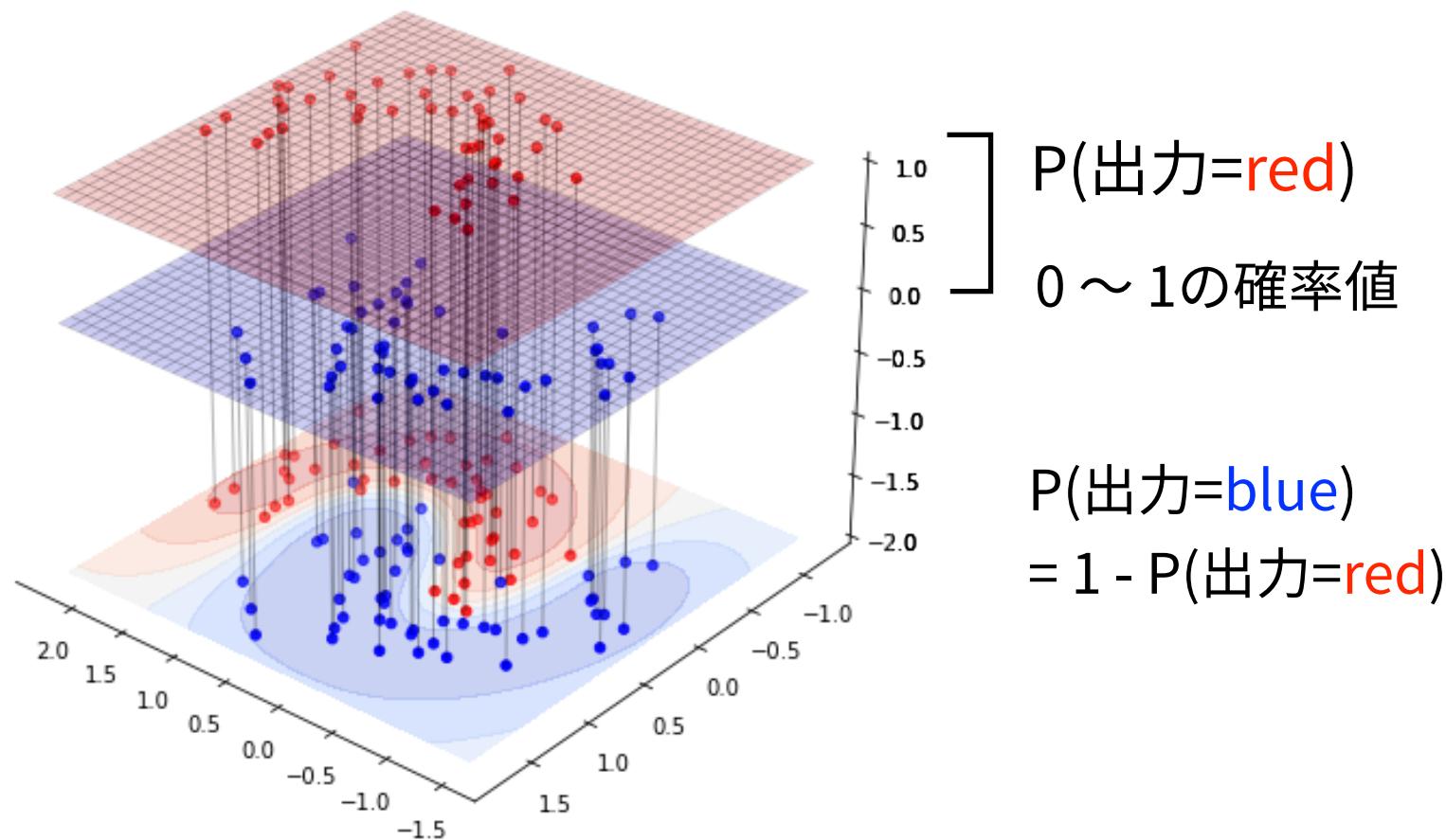
→ 境界線の引き方の方針



# 機械学習は「新しいコンピュータプログラムの作り方」

内部原理は「曲面モデル」を点にフィッティングしているだけ

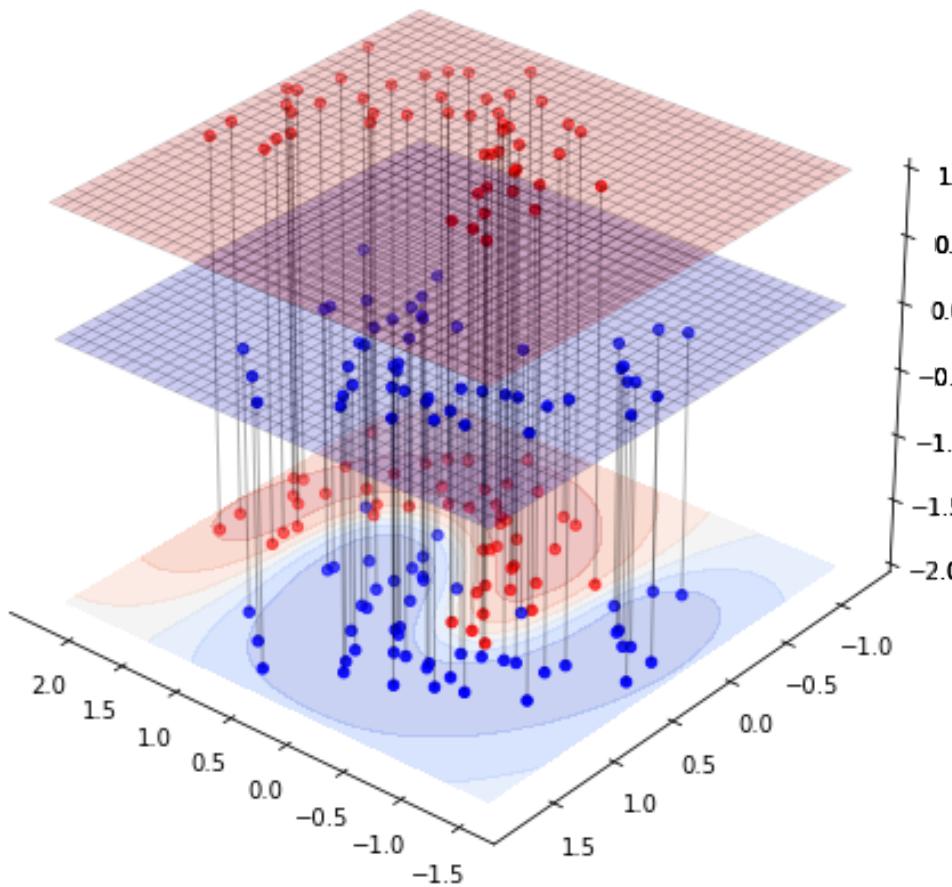
→ 境界線の引き方の方針



# 機械学習は「新しいコンピュータプログラムの作り方」

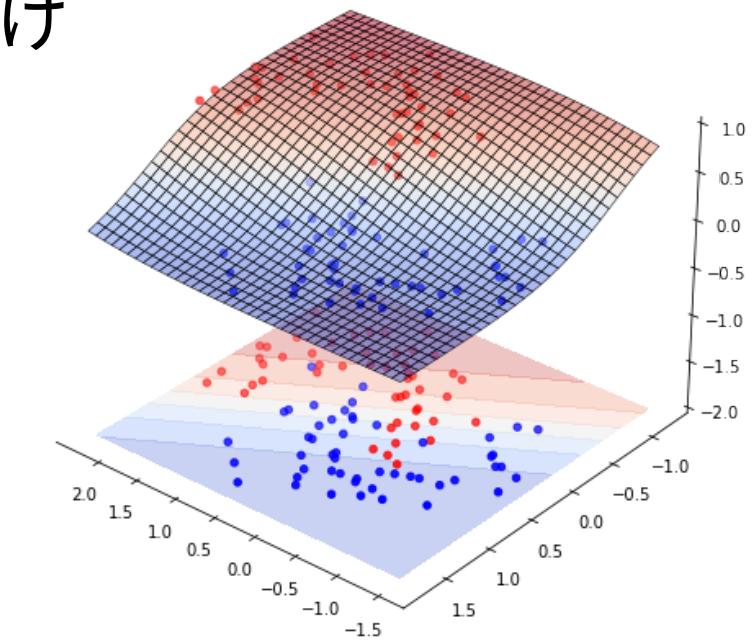
内部原理は「曲面モデル」を点にフィッティングしているだけ

→ 境界線の引き方の方針

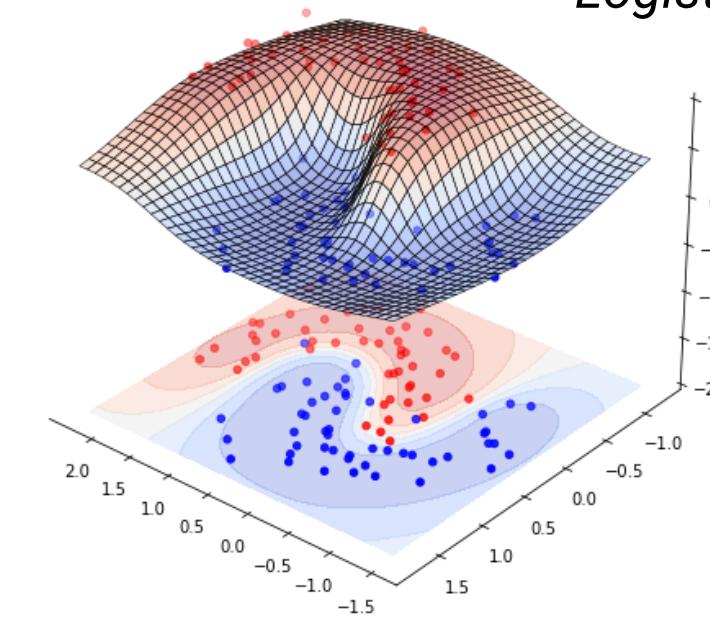


$P(\text{出力}=\text{red})$   
0 ~ 1 の確率値

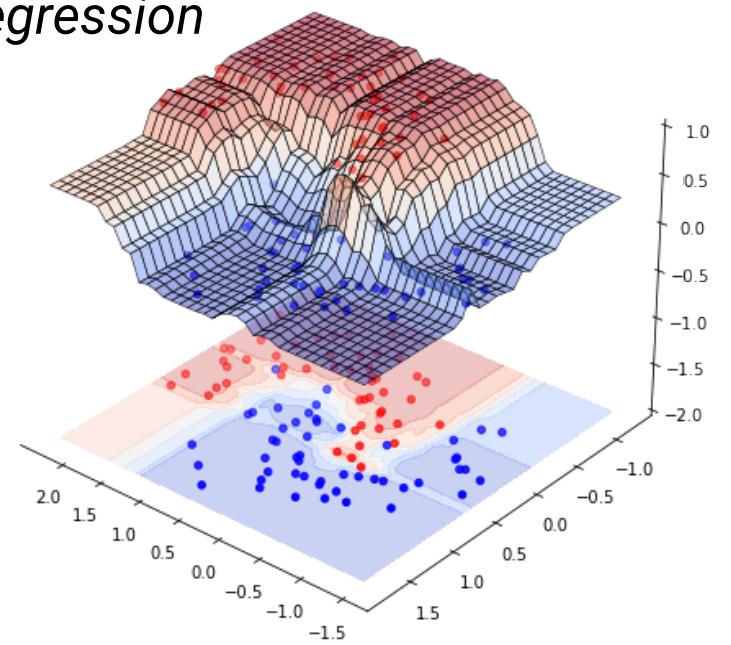
$P(\text{出力}=\text{blue})$   
 $= 1 - P(\text{出力}=\text{red})$



Logistic Regression

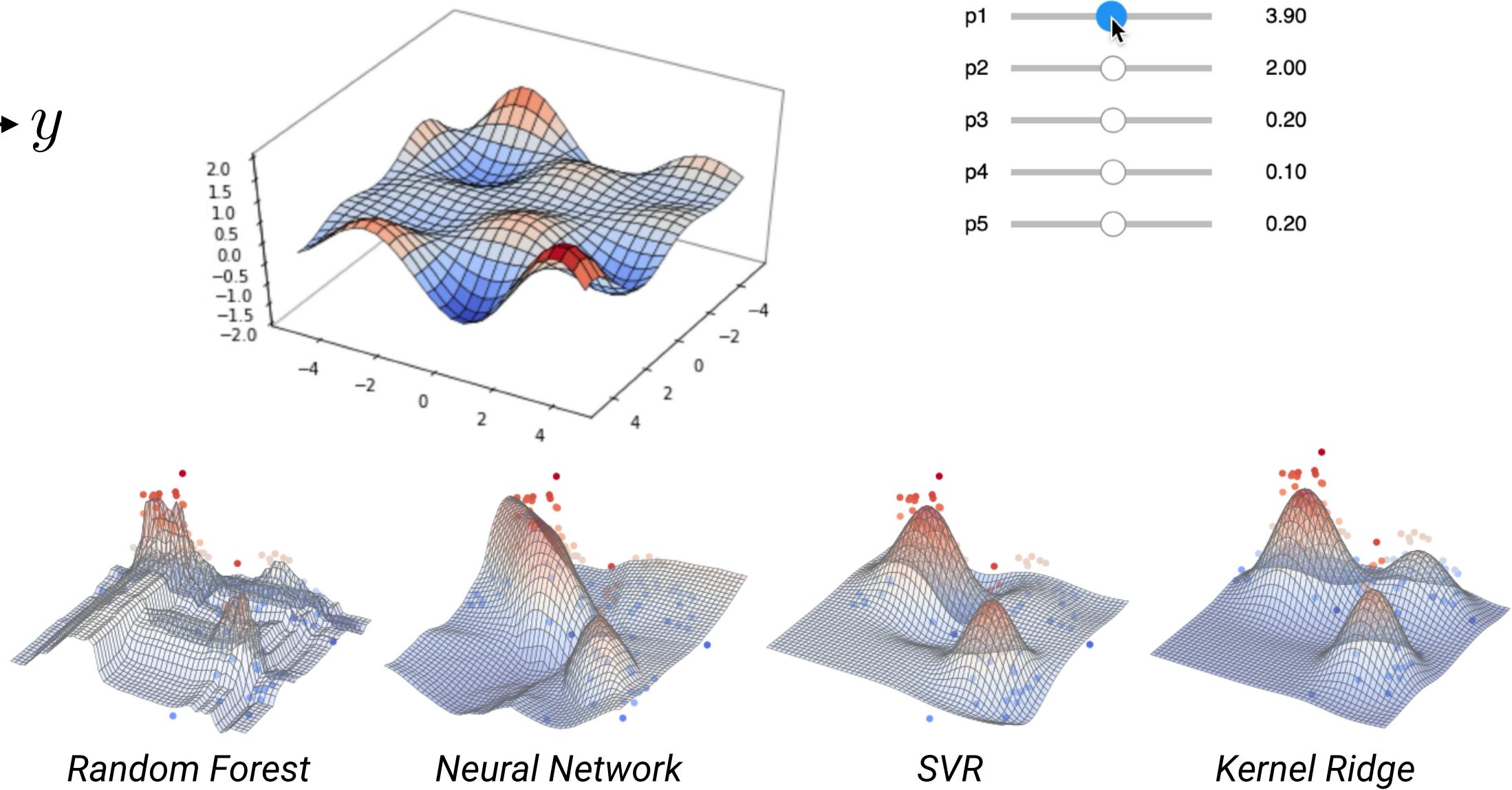
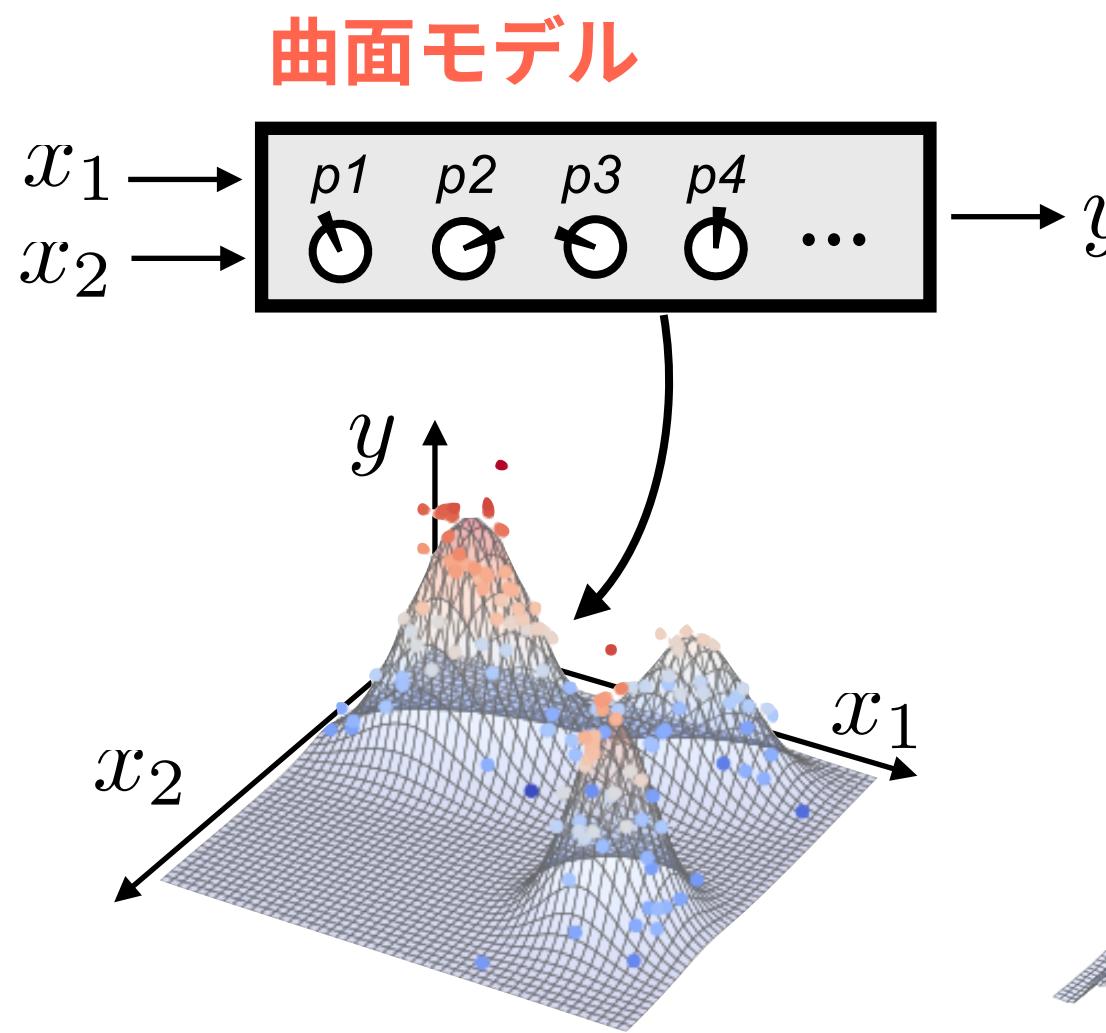


Gaussian Process



Random Forest

「曲面モデル」の内部パラメタ値を調整して見本点にあうようフィッティングする



# 蛇足：希望的呼び方による幻想にご注意を



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

*wishful mnemonics*

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

<https://arxiv.org/abs/2104.12871>

Computer Science > Artificial Intelligence

[Submitted on 26 Apr 2021 ([v1](#)), last revised 28 Apr 2021 (this version, v2)]

## Why AI is Harder Than We Think

Melanie Mitchell

Since its beginning in the 1950s, the field of artificial intelligence has cycled several times between periods of optimistic predictions and massive investment ("AI spring") and periods of disappointment, loss of confidence, and reduced funding ("AI winter"). Even with today's seemingly fast pace of AI breakthroughs, the development of long-promised technologies such as self-driving cars, housekeeping robots, and conversational companions has turned out to be much harder than many people expected. One reason for these repeating cycles is our limited understanding of the nature and complexity of intelligence itself. In this paper I describe four fallacies in common assumptions made by AI researchers, which can lead to overconfident predictions about the field. I conclude by discussing the open questions spurred by these fallacies, including the age-old challenge of imbuing machines with humanlike common sense.

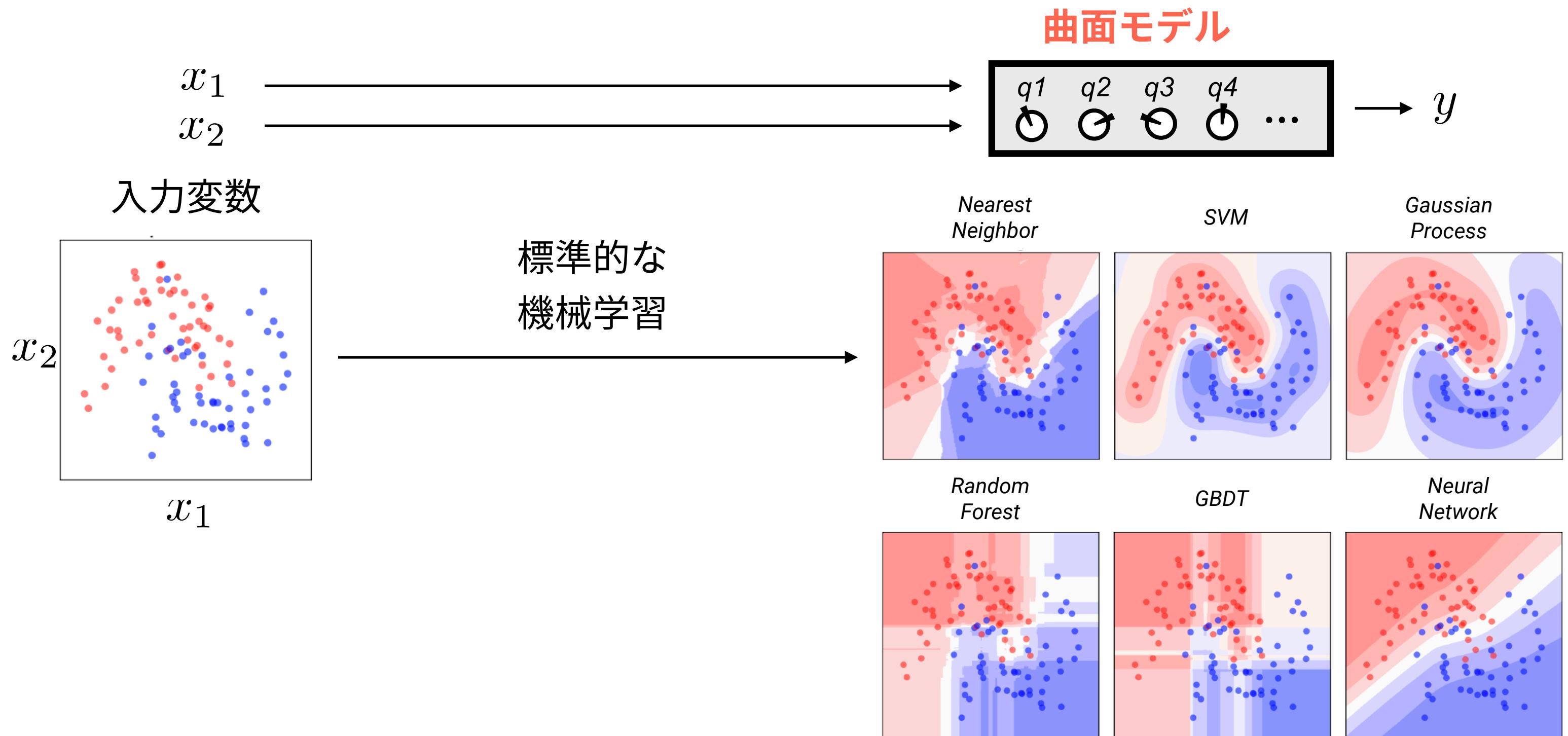
SIGART Newsletter No. 57 April 1976

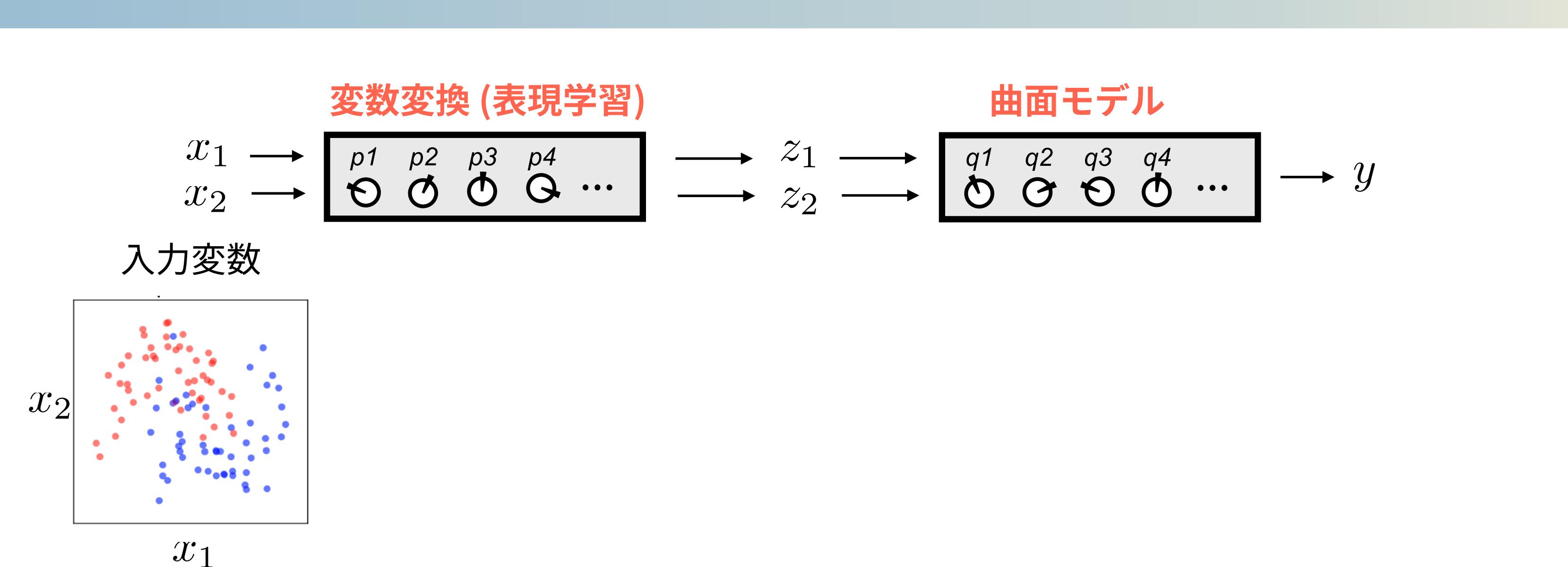
## ARTIFICIAL INTELLIGENCE MEETS NATURAL STUPIDITY

Drew McDermott

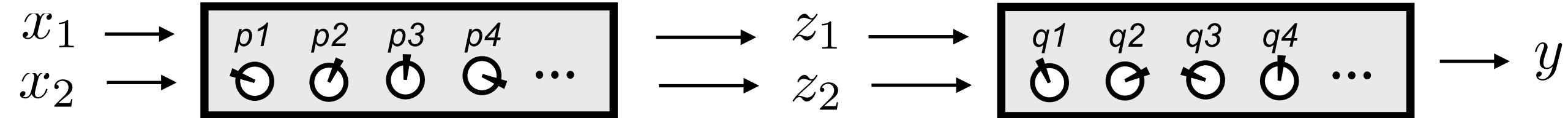
MIT AI Lab Cambridge, Mass 02139

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.

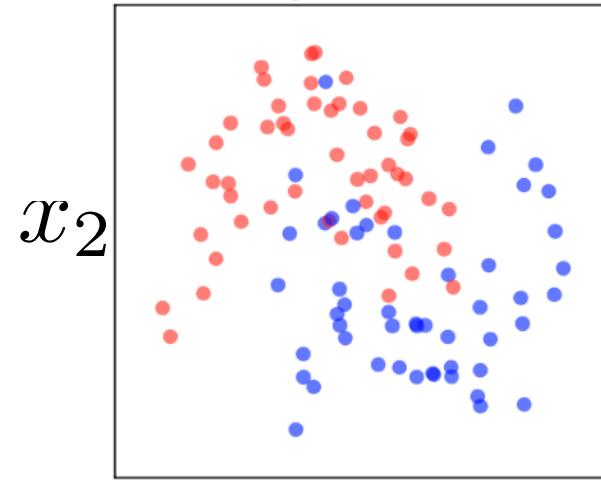




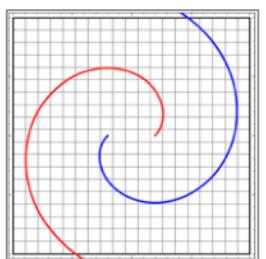
## 変数変換(表現学習)



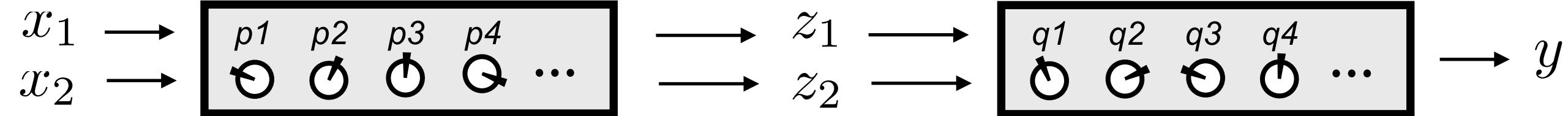
入力変数



## 曲面モデル

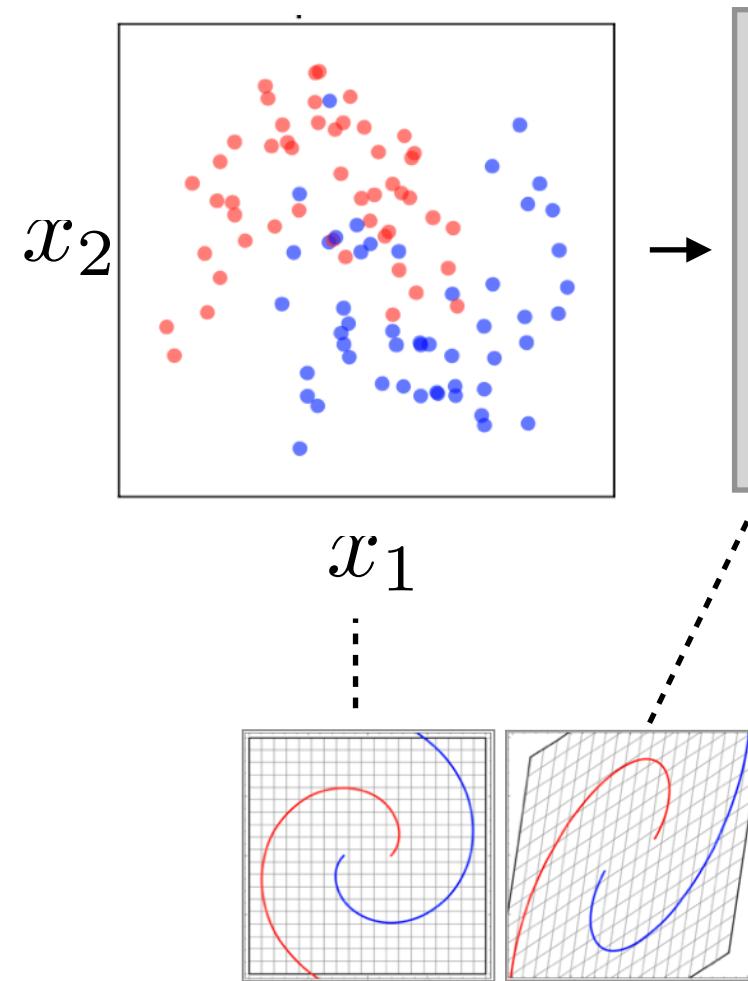
 $x_1$ 

## 変数変換(表現学習)

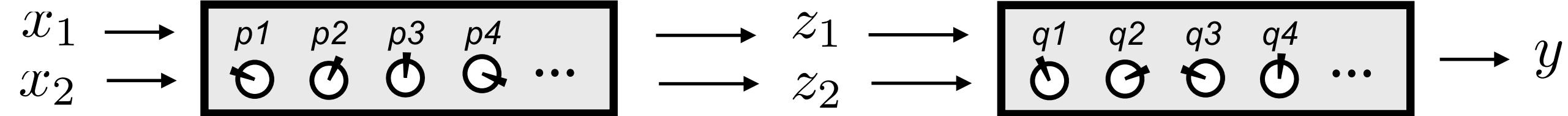


## 曲面モデル

入力変数

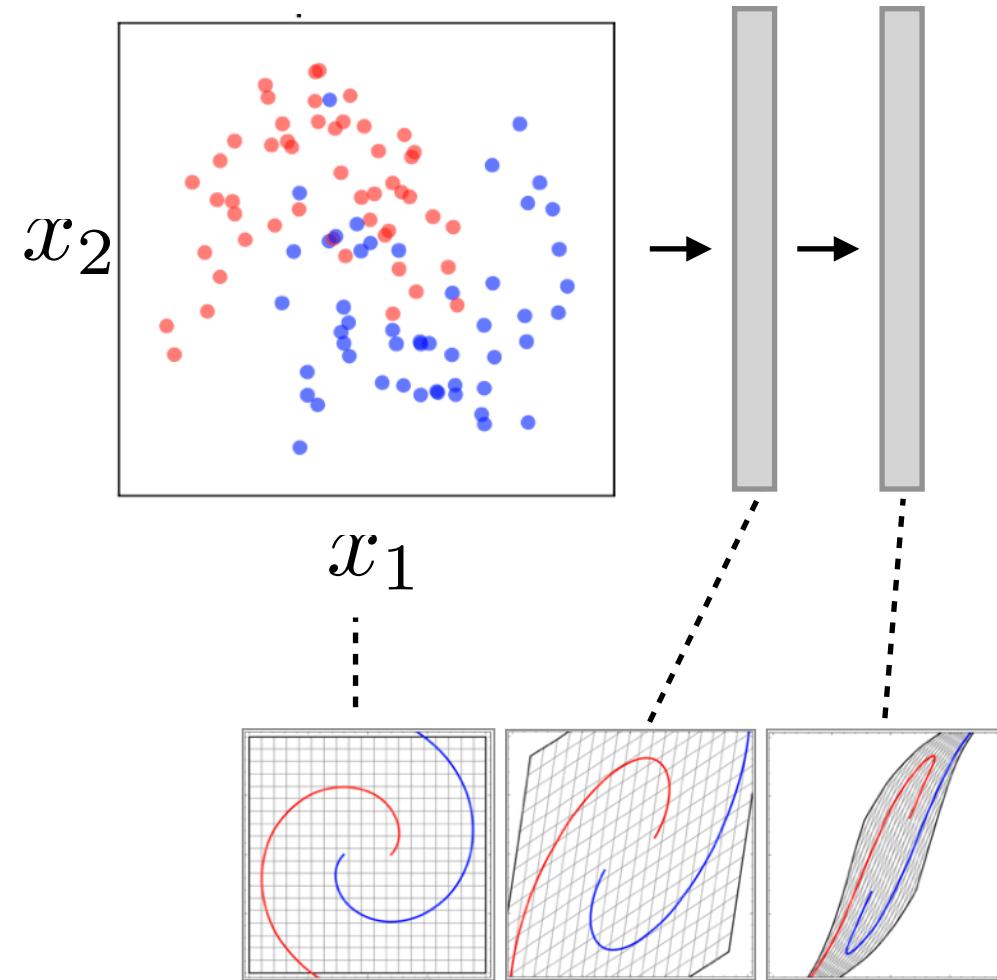


## 変数変換(表現学習)

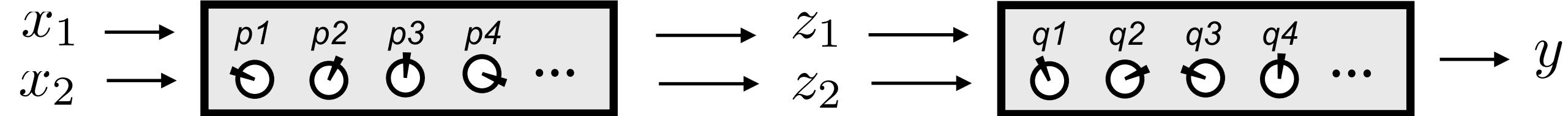


## 曲面モデル

## 入力変数

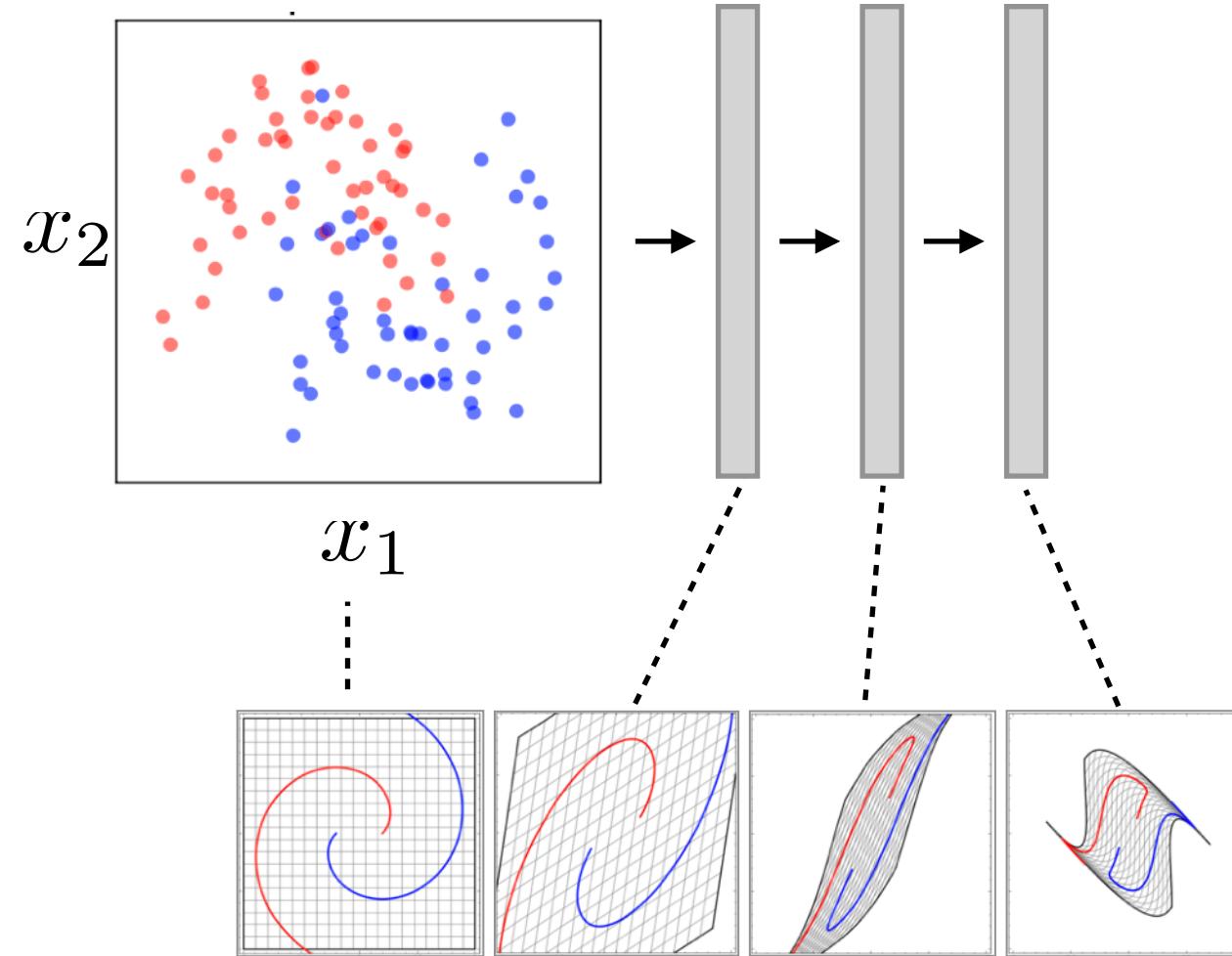


## 変数変換(表現学習)

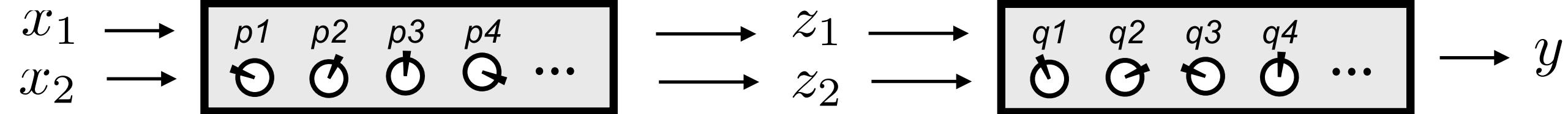


## 曲面モデル

## 入力変数

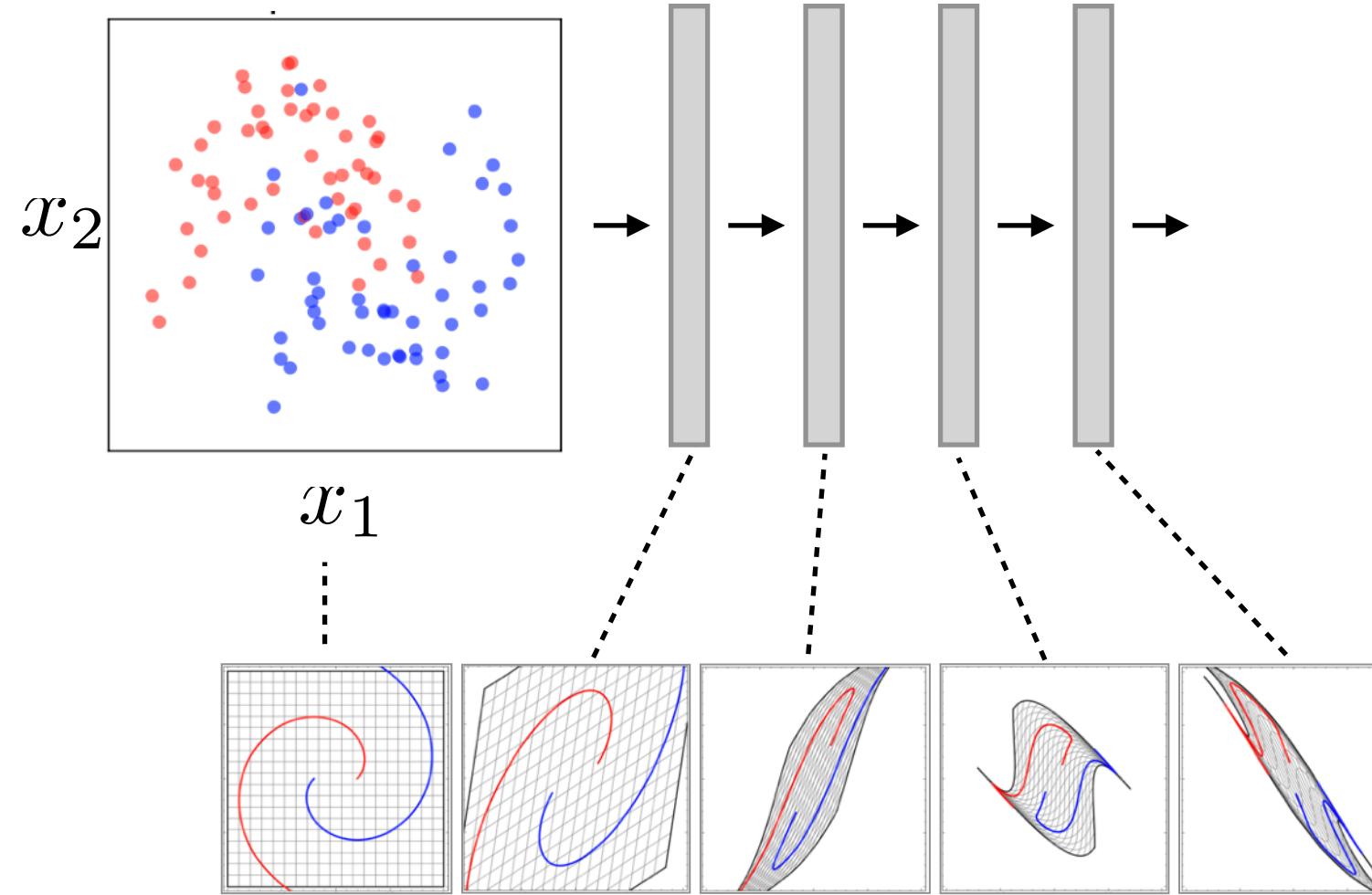


## 変数変換(表現学習)

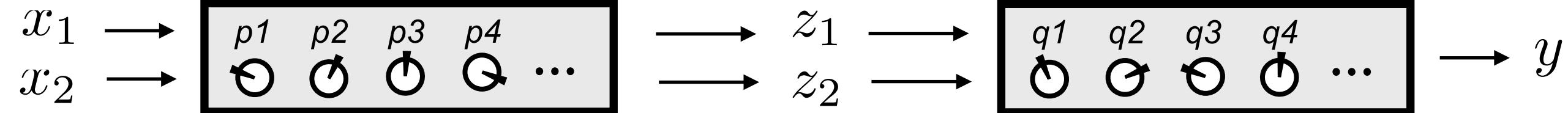


## 曲面モデル

## 入力変数

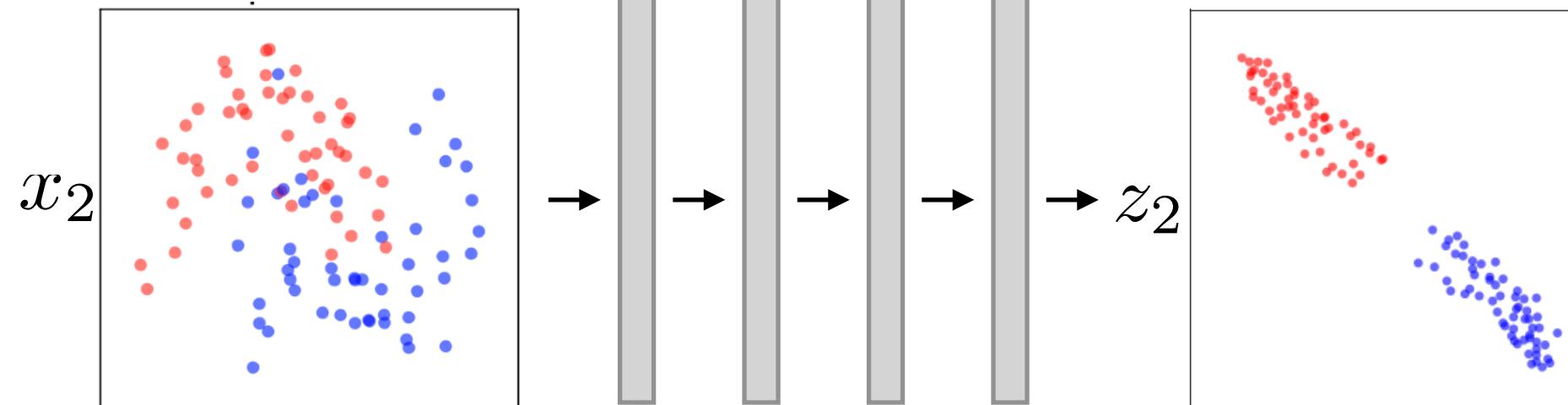


## 変数変換(表現学習)

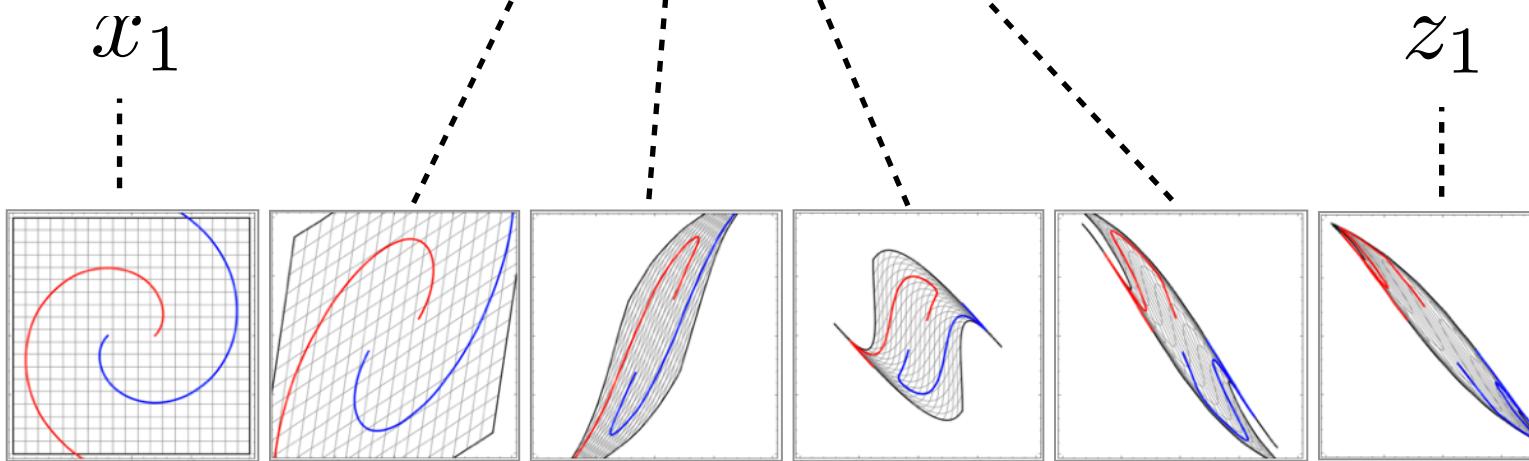


## 曲面モデル

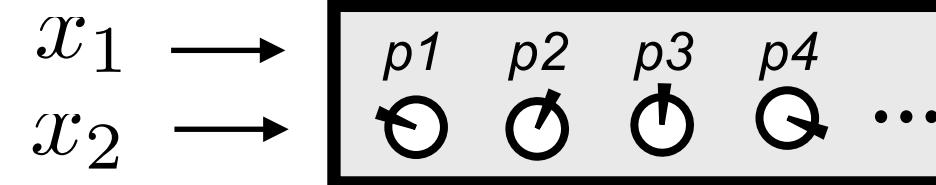
入力変数



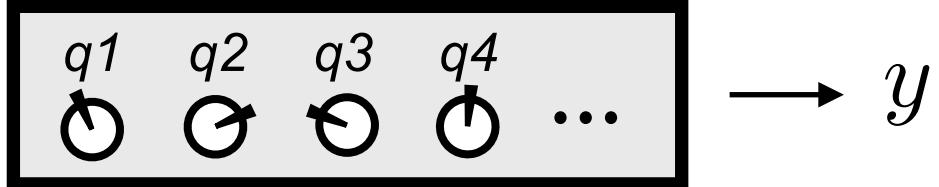
潜在変数



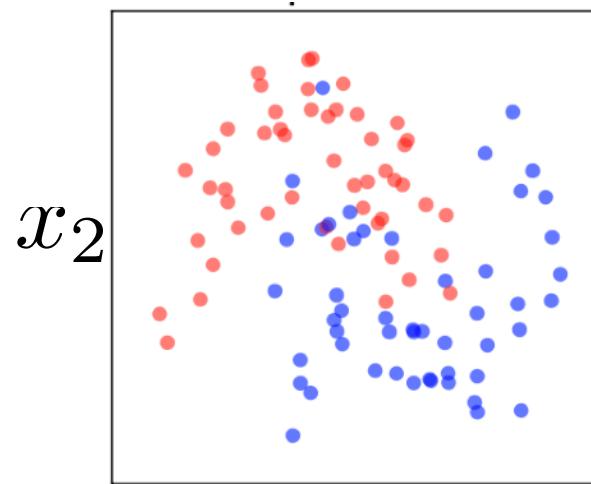
## 変数変換(表現学習)



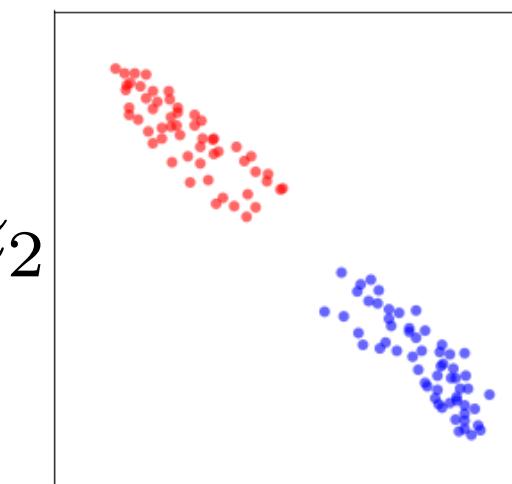
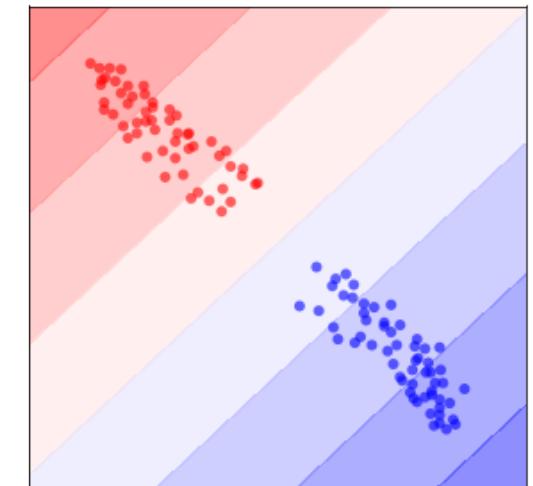
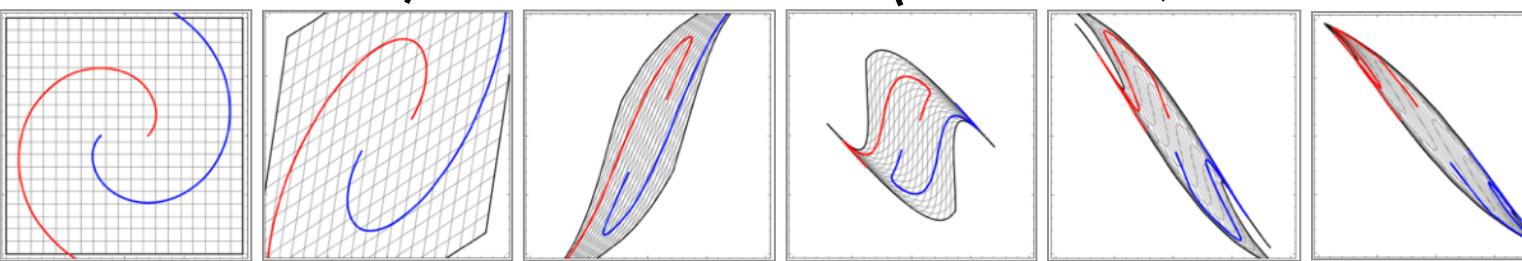
## 曲面モデル



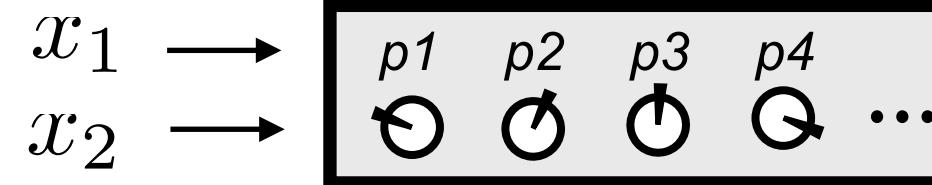
入力変数



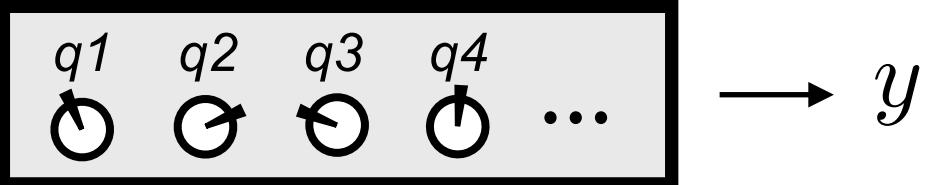
潜在変数

標準的な  
機械学習 $x_1$  $z_1$ 良い変数さえ見つかれば  
ここはシンプルで良い！

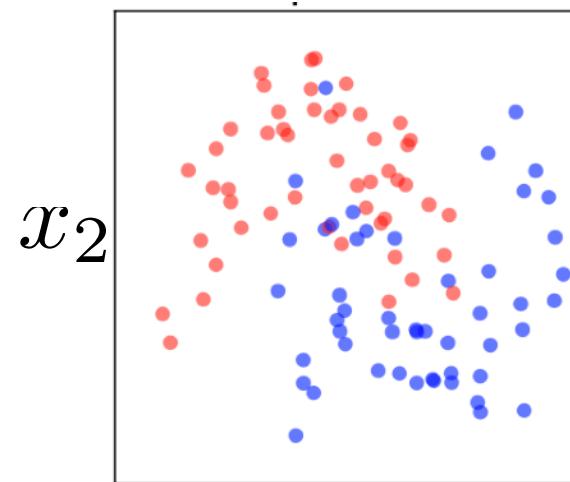
## 変数変換(表現学習)



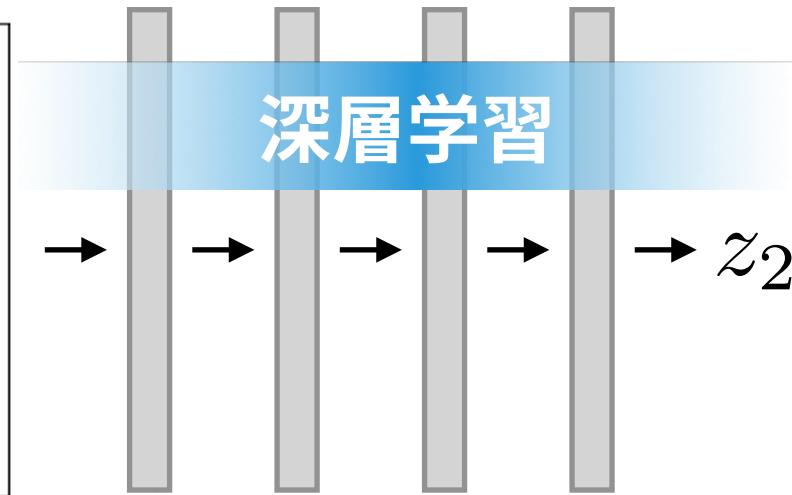
## 曲面モデル



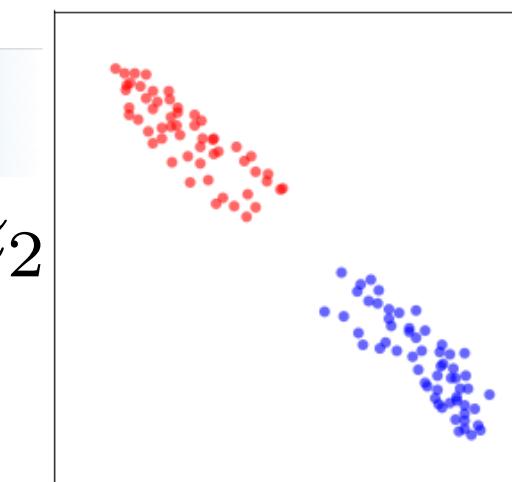
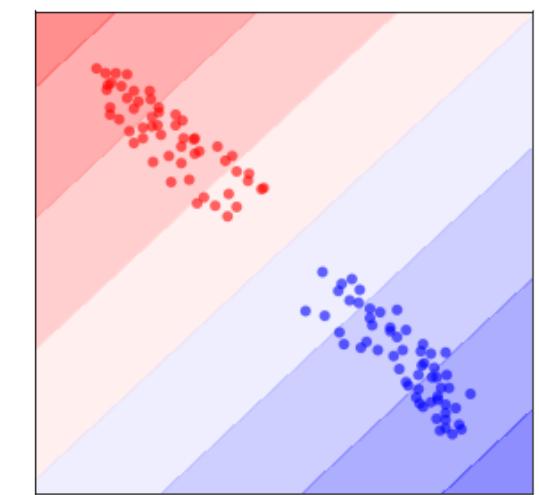
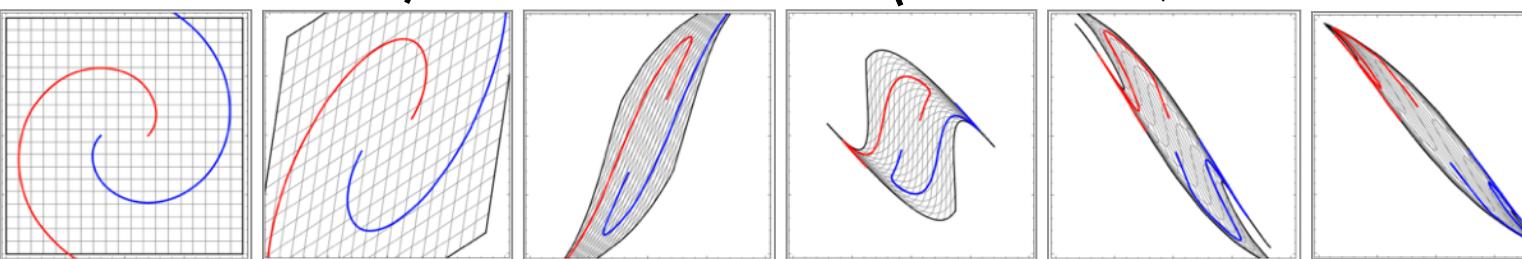
入力変数

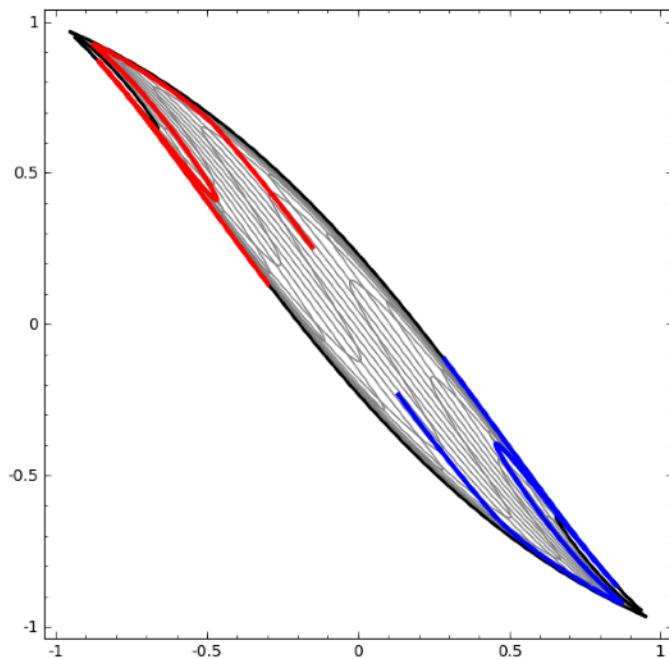
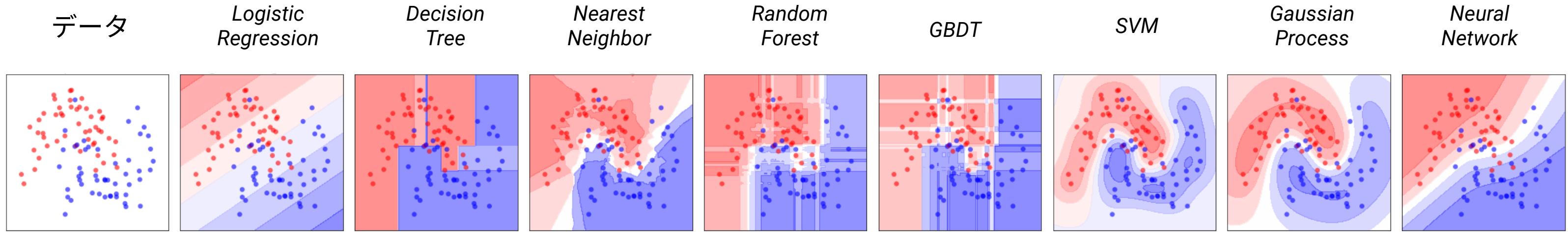


深層学習

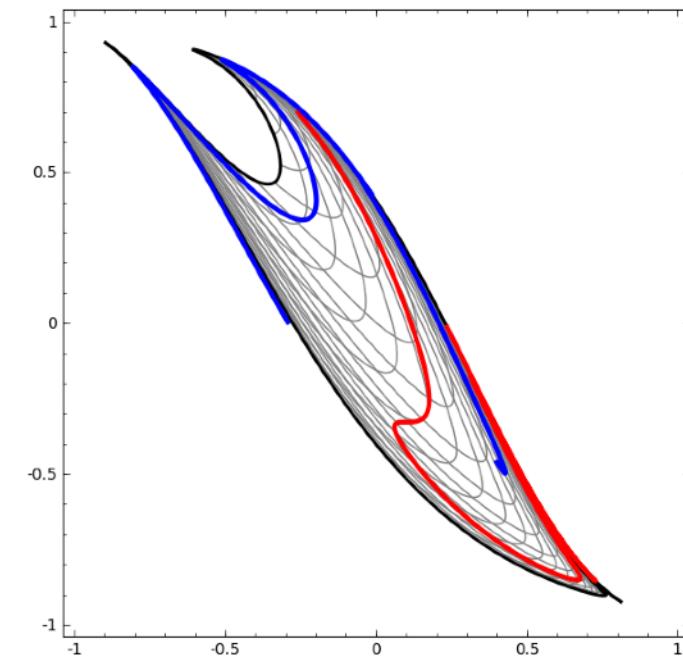


潜在変数

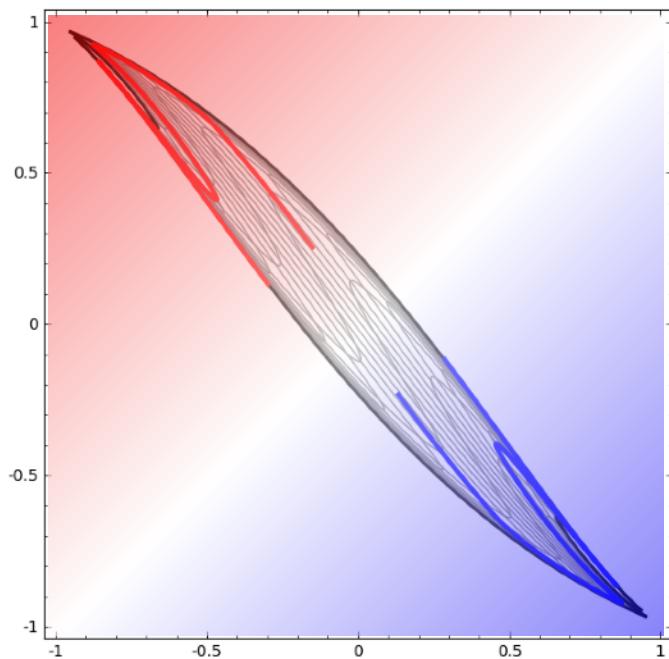
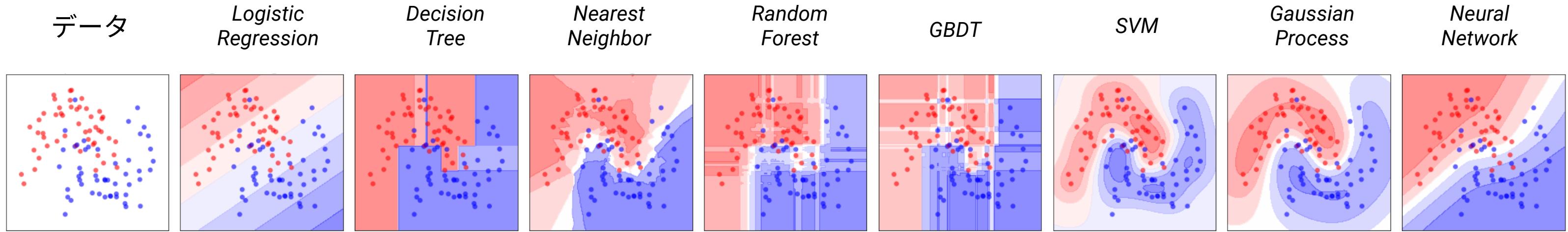
標準的な  
機械学習 $x_1$  $z_1$ 良い変数さえ見つかれば  
ここはシンプルで良い！



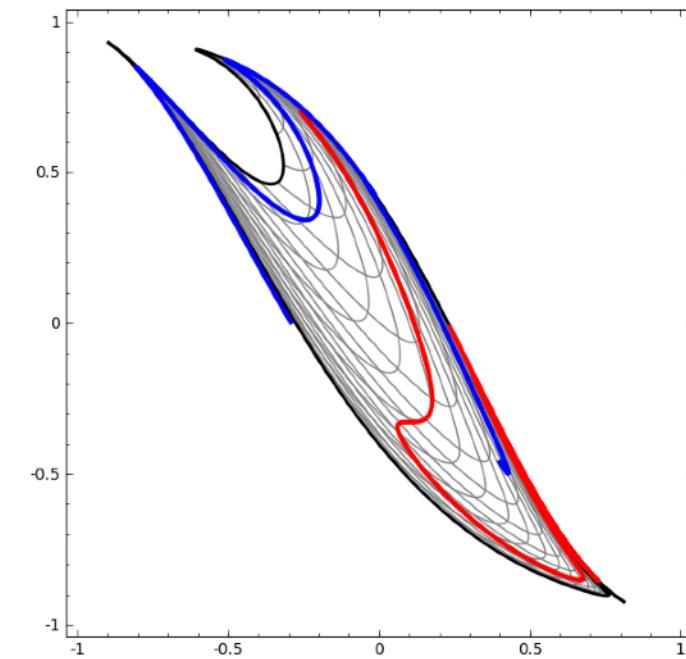
できるだけ線で  
分けられるように  
変換を学習



いつもうまくいく  
とは限らない…。  
間違えると元より  
難しい問題になる！



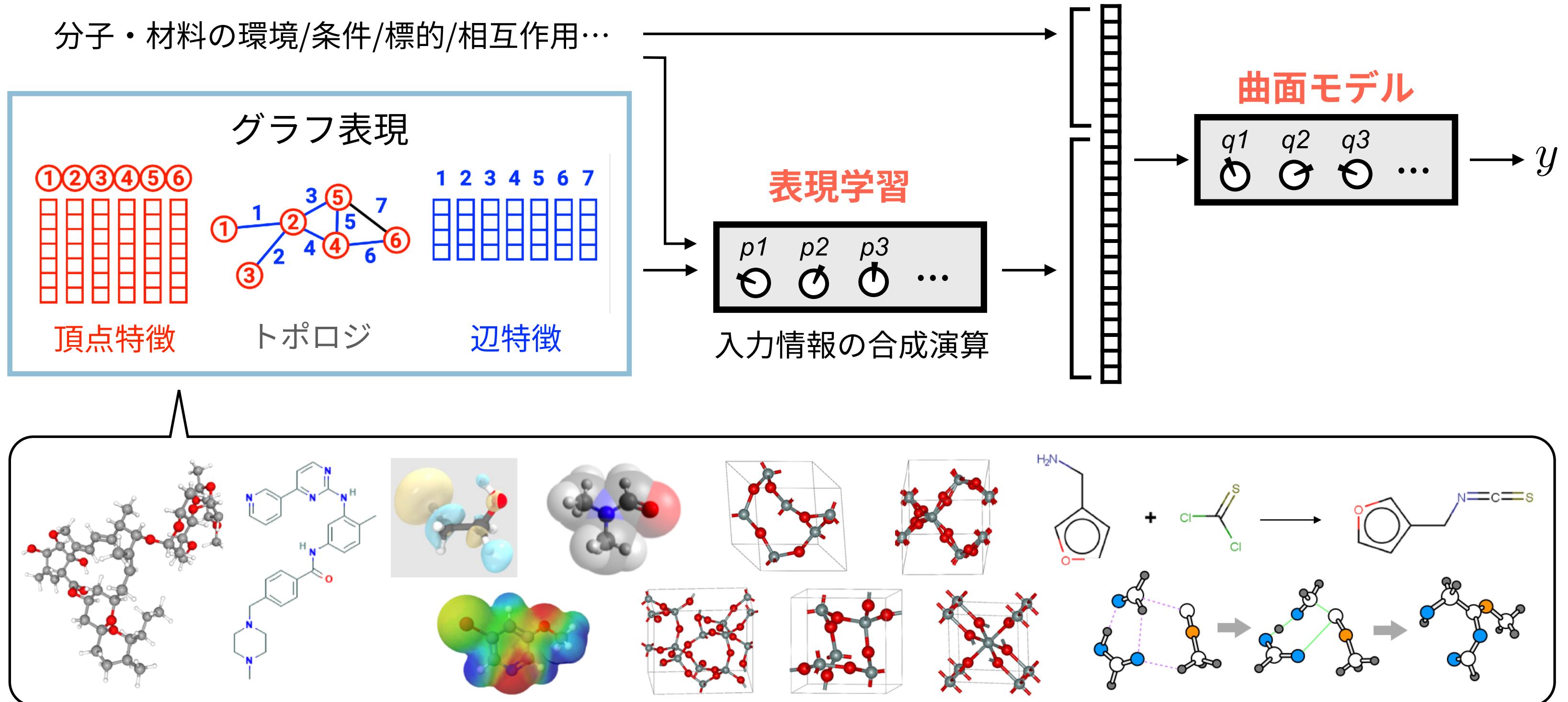
できるだけ線で  
分けられるように  
変換を学習



いつもうまくいく  
とは限らない…。  
間違えると元より  
難しい問題になる！

# 実例：分子・材料の表現学習とGraph Neural Networks

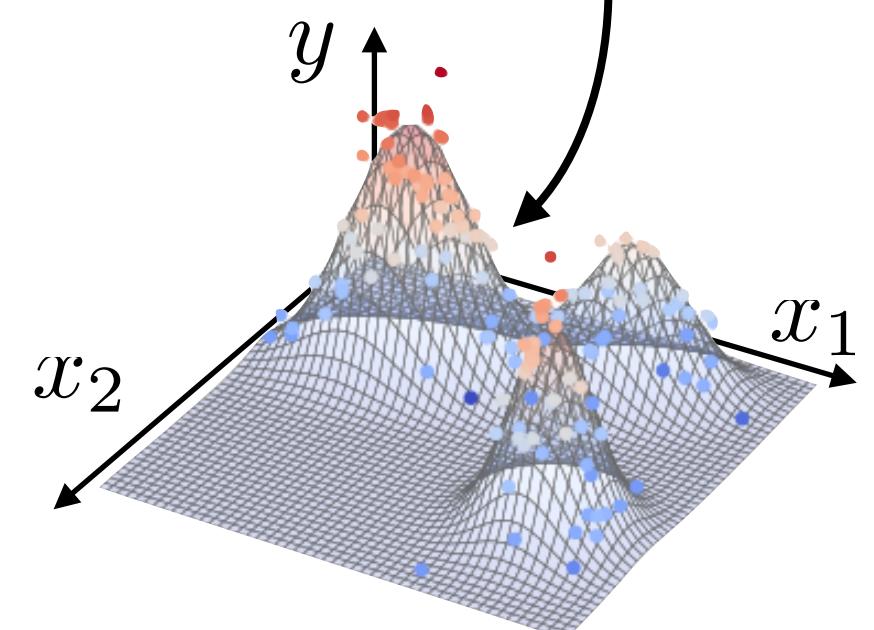
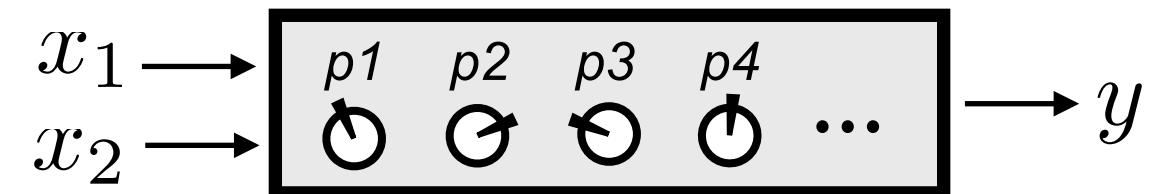
分子・材料の環境/条件/標的/相互作用…



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

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

## 曲面モデル

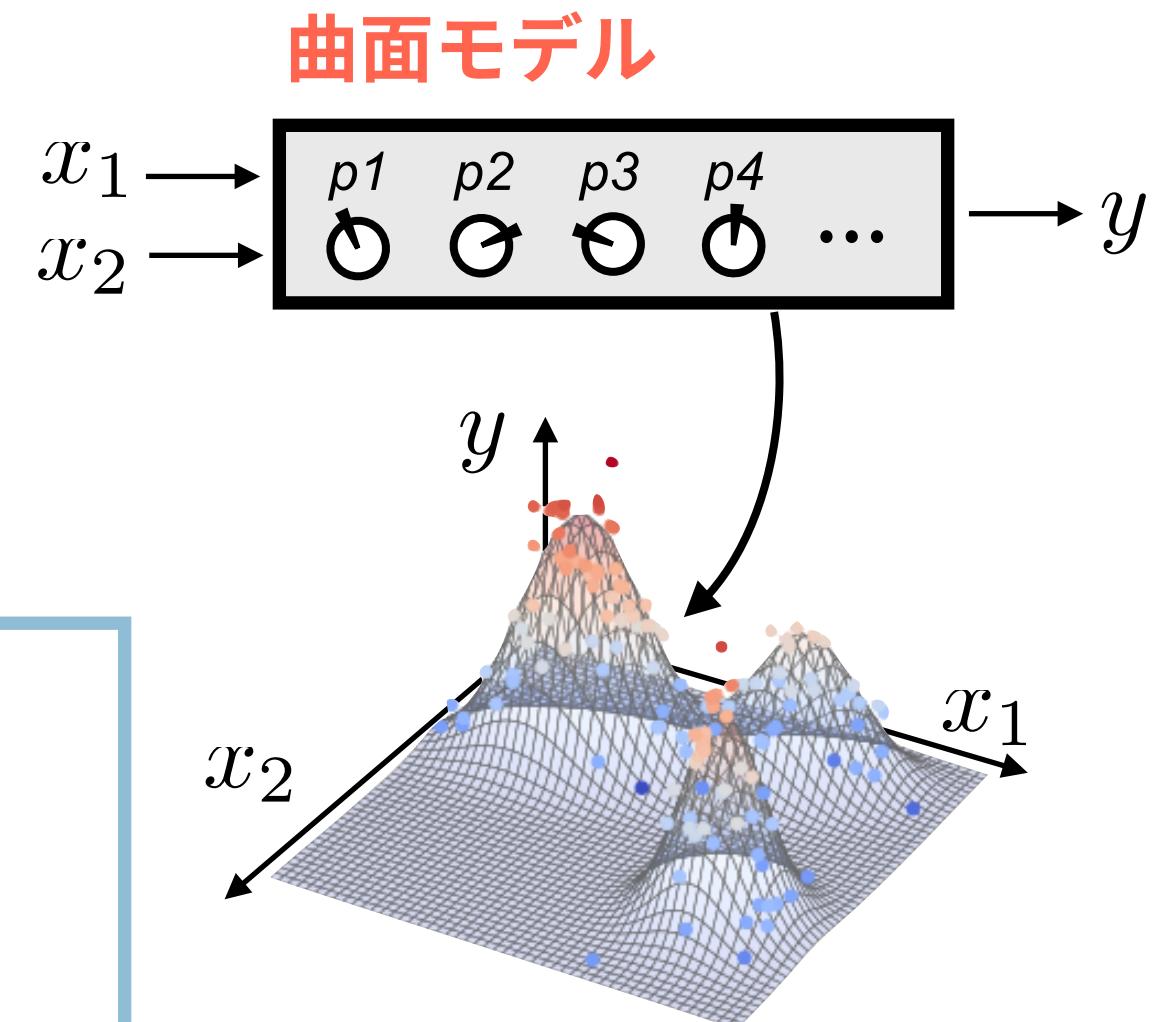


右の絵を眺めると、機械学習が  $x_1$  と  $x_2$  以外の  
入力されてない情報を全く考慮してくれないことは一目瞭然

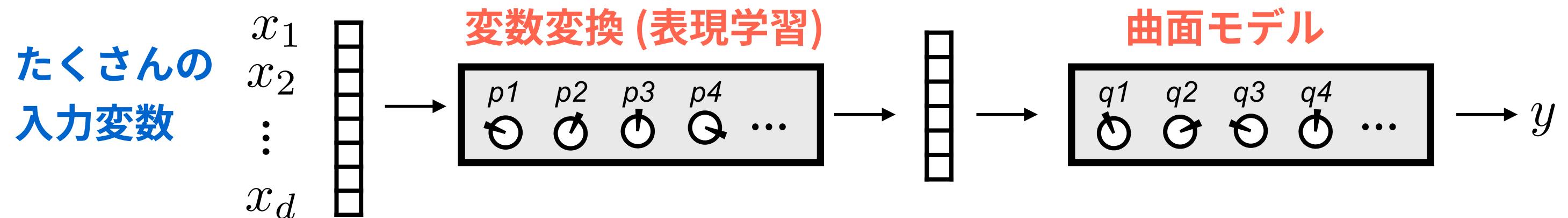
→ 出力の予測に本当は必要な情報を入力していなかつたら  
機械学習は擬似相関に過ぎず何も本質を捉えられない  
*spurious correlation*

## 機械学習×化学：スタートラインでのつまづき

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

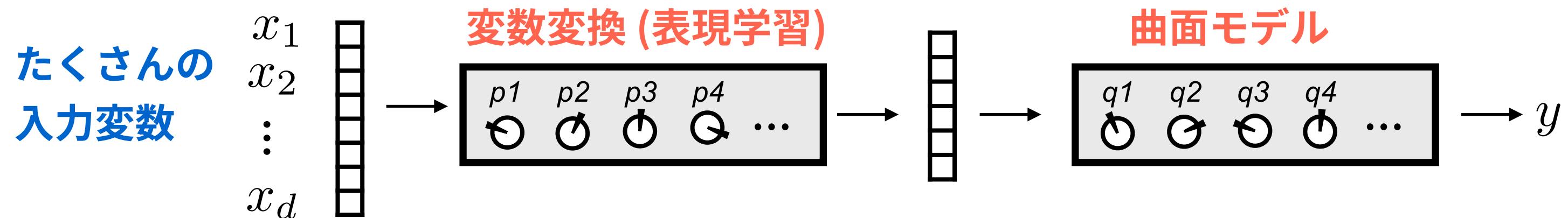


必然的に「少しでも関係ありそうな情報は入力して」表現学習で重要なものを峻別という戦略に

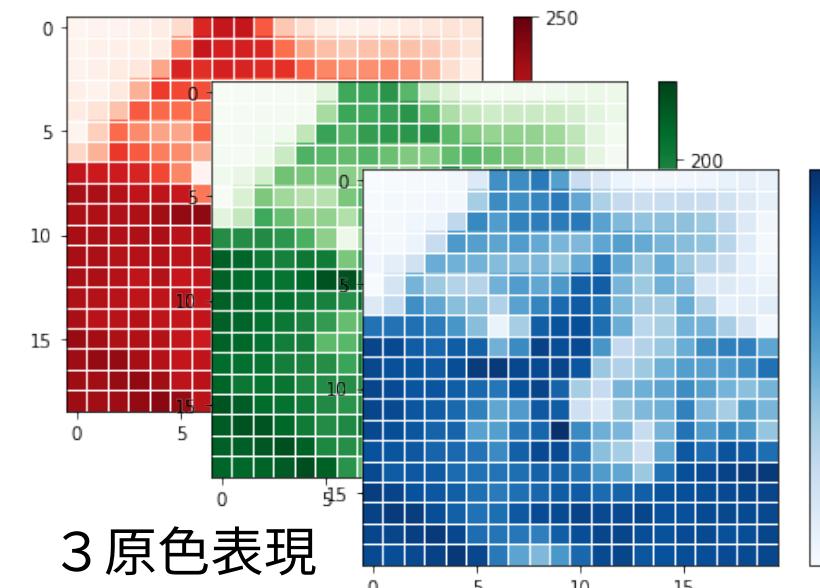
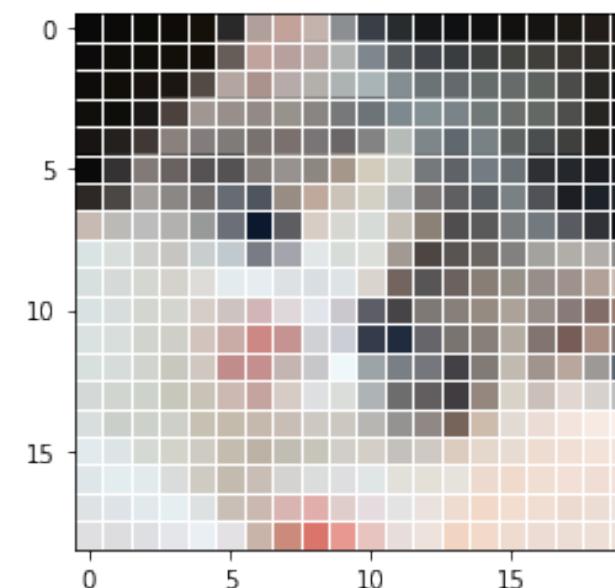


# 機械学習の現代的側面① 入力変数がめっちゃ多い…

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



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

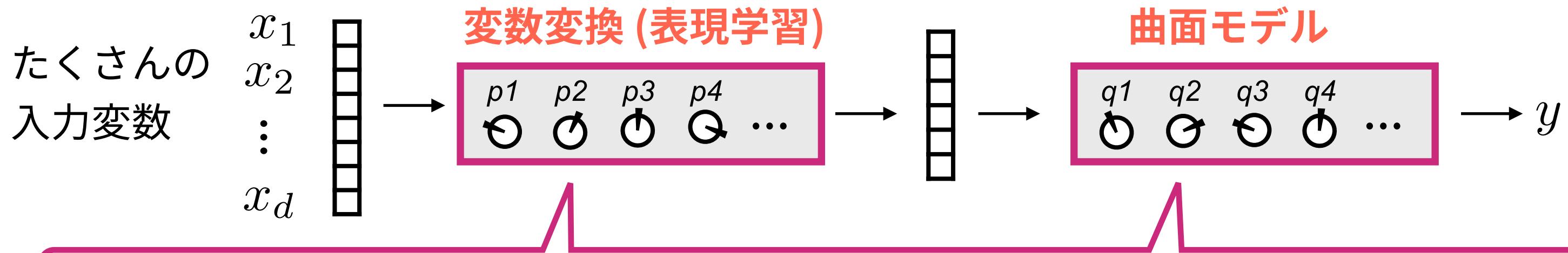


3原色表現

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

20×20 → 1,200変数  
1000×1000 → 300,000変数

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



ResNet50: 2600万パラメタ

ResNet101: 4500万パラメタ

EfficientNet-B7: 6600万パラメタ

VGG19: 1億4400万パラメタ

12-layer, 12-heads BERT: 1億1000万パラメタ

24-layer, 16-heads BERT: 3億3600万パラメタ

GPT-2 XL: 15億5800万パラメタ

GPT-3: 1750億パラメタ

現代の機械学習の技術研究が向き合う設定：

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

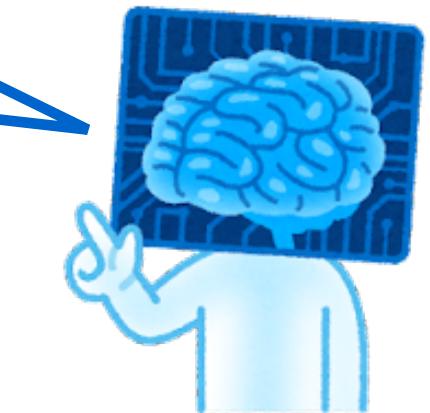
現代の機械学習は強力で**良いデータが十分にあれば複雑な入出力関係でも学習できる！**

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

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

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

by ディープラーニング様



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

**羅生門効果：良い機械学習モデルの多重性（非一意性）**

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

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

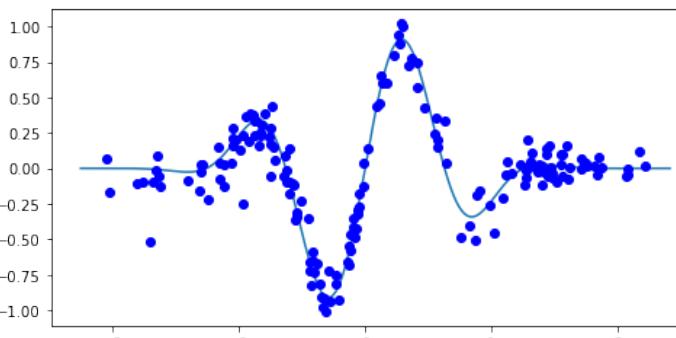
## 羅生門効果：良い機械学習モデルの多重性（非一意性）

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

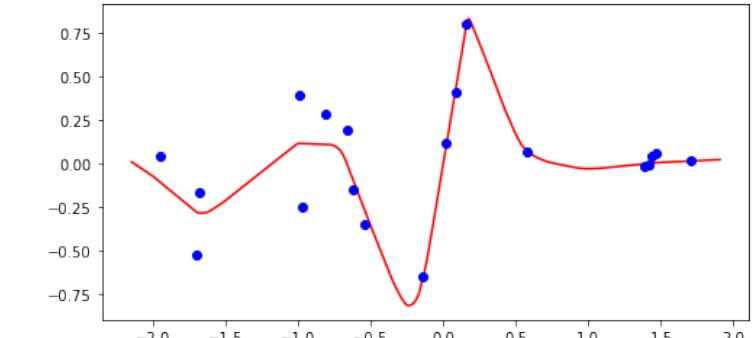
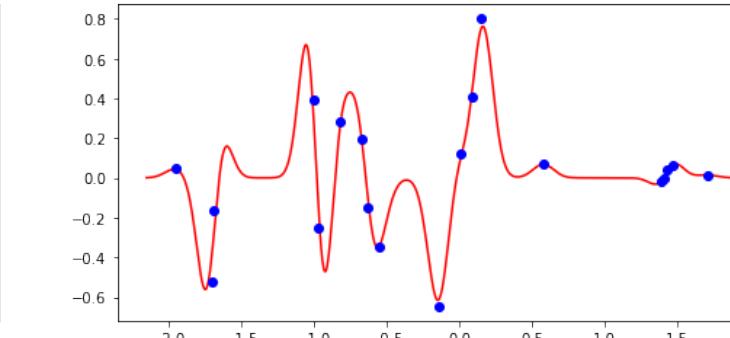
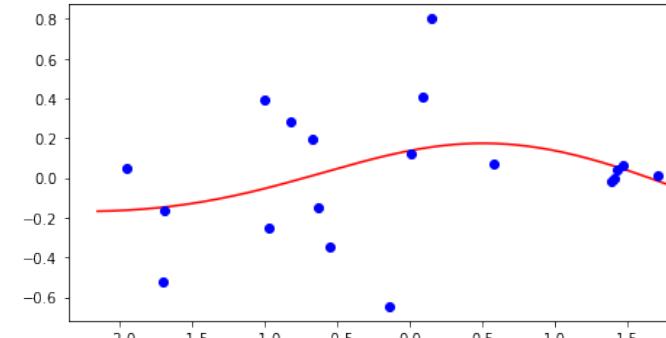
原因：入力変数の選び方、モデルの選択や設計、初期値の違い、パラメタの多重性、…

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

だいたいの方法で類似



手法やモデルによって予測時の挙動にかなり差が出てしまう



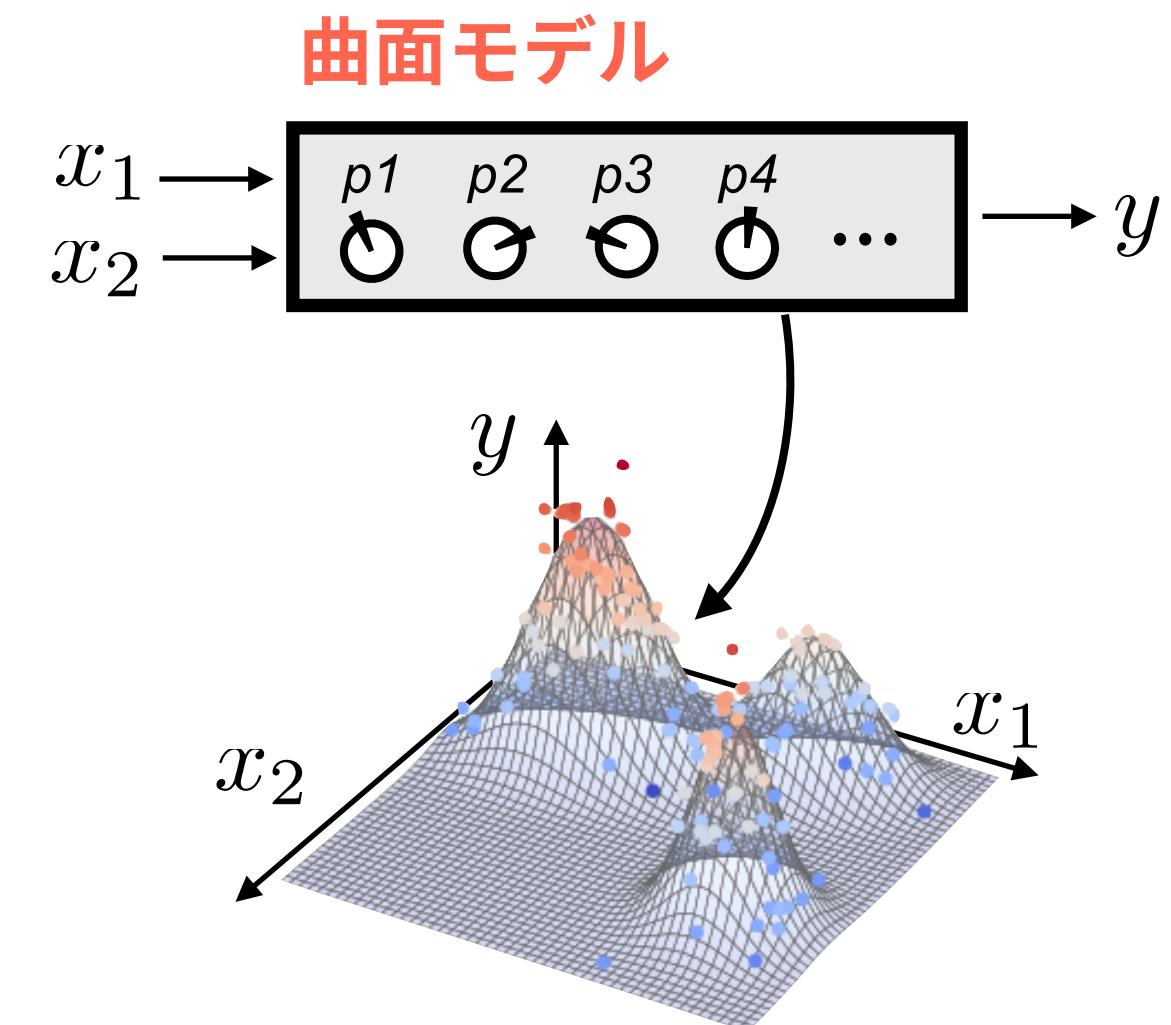
# 現代の機械学習の技術的関心はこの高次元性をどう手懐けるか



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

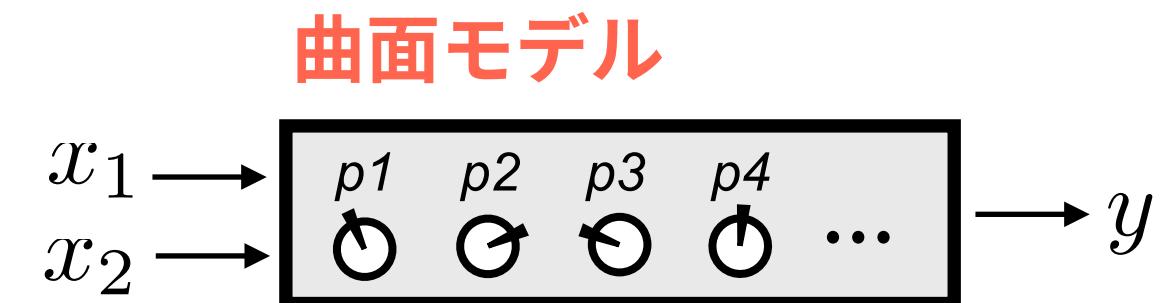
1. 確率的最適化・正則化 → モデルが大きい自由度の中で暴れまくらないよう動ける範囲を何とかして制御・制限・安定化する
2. 事前学習 (Warm Start) の転移 → 事前に探しておいた良い感じのパラメタ初期値を使う
3. 帰納バイアスの設計

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



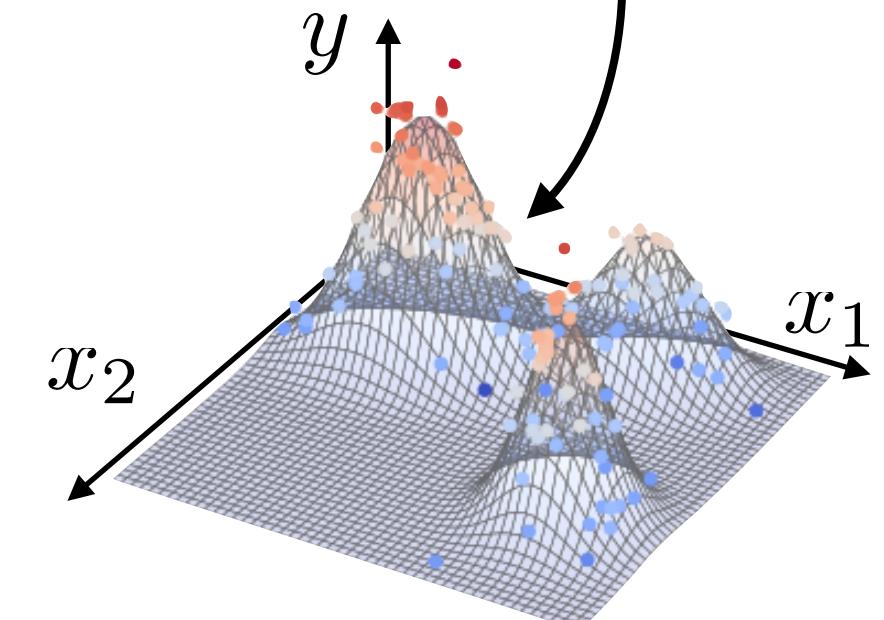
1. 確率的最適化・正則化 → モデルが大きい自由度の中で暴れまくらないよう動ける範囲を何とかして制御・制限・安定化する
2. 事前学習 (Warm Start) の転移 → 事前に探しておいた良い感じのパラメタ初期値を使う
3. 帰納バイアスの設計

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



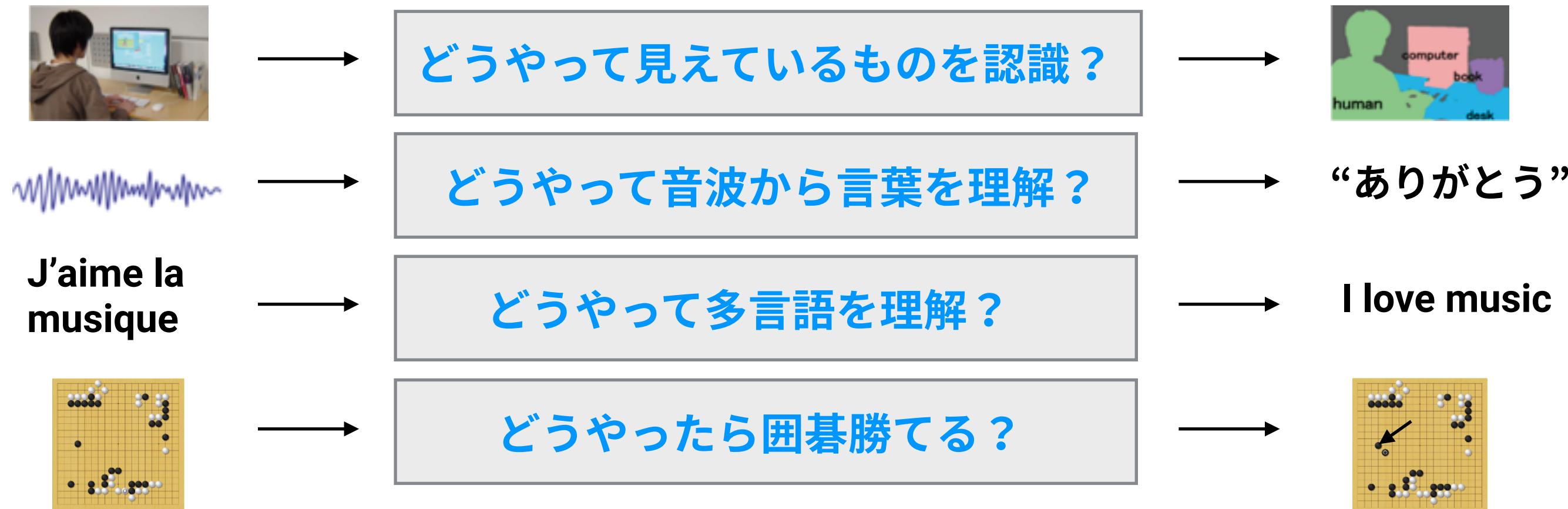
## 機械学習×化学：化学に適合した帰納バイアスのデザイン

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



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

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



# 因果の理解には実験研究(介入研究)が必要不可欠

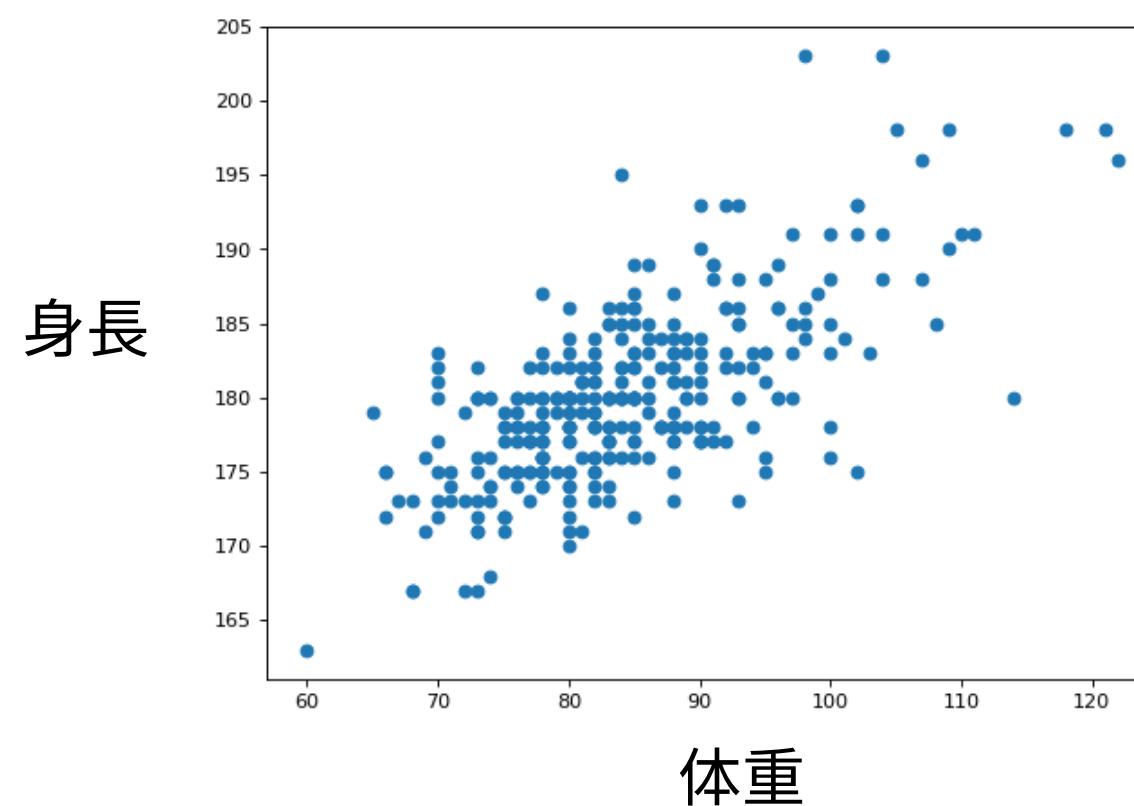


機械学習はあくまでデータの中の多次元相関を捉え、それによって予測する技術

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

## 日本プロ野球開幕一軍選手の身長・体重データ

(2016年球団公式サイト選手データより自作)



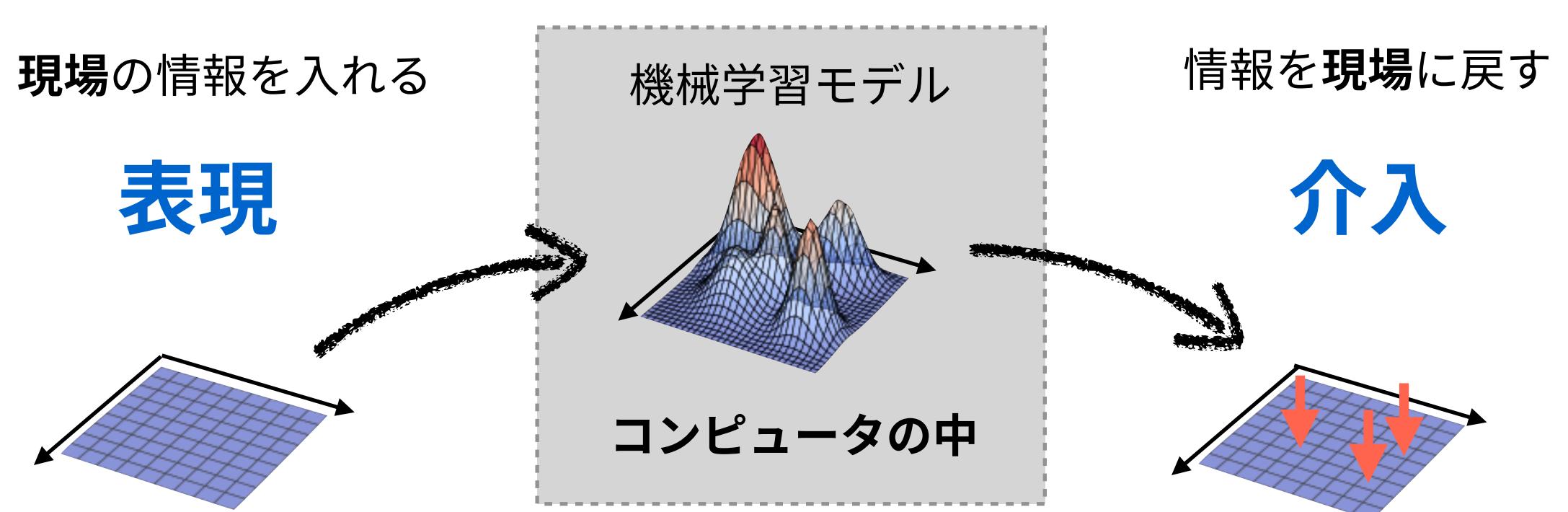
「体重を増やせば身長も伸びる」が正しいかは  
この観察データだけからは決して分からぬ

## 応用統計学の基本のキ

相関関係は必ずしも因果関係を意味しない

→ 「予測ができる」ことは「理解」や「発見」ができる<sup>ことを直接は意味しない！！</sup>

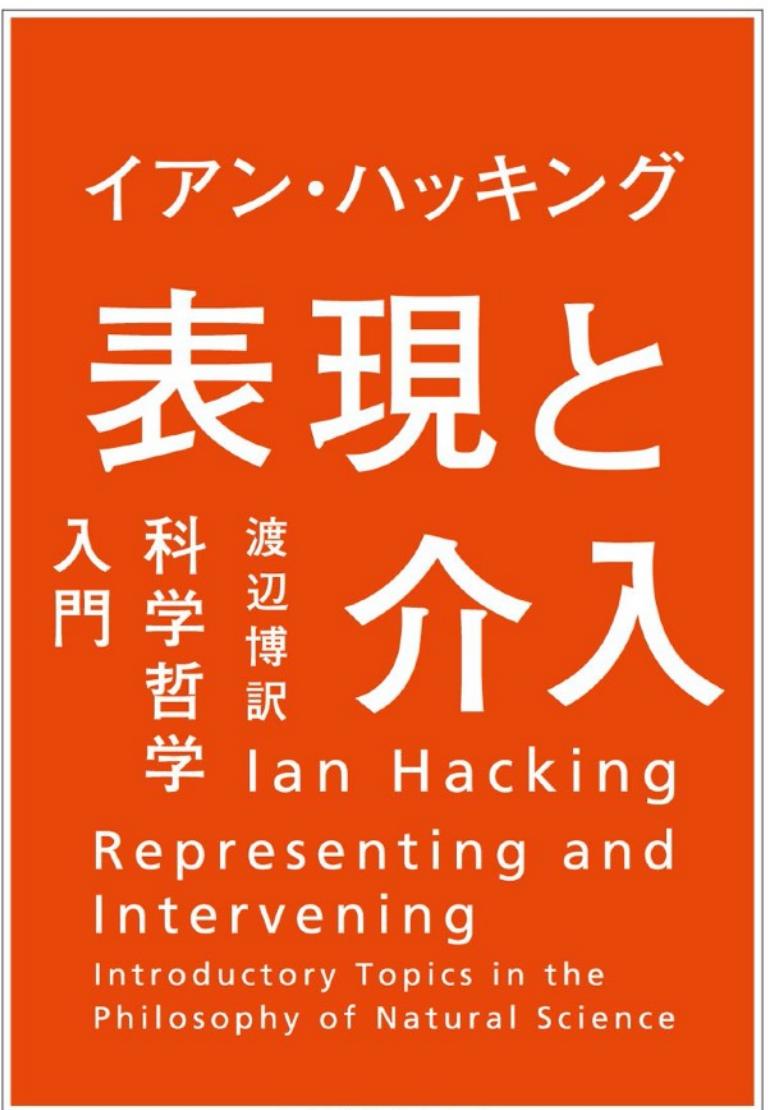
事件はコンピュータ(機械学習)の中で起きてるんじゃない、**現場**で起きているんだ！ by 俺



- 有効な入力変数の同定
- 機械学習タスクのデザイン
- 訓練データの設計と収集
- 既存のデータ・知識の利用

**現場**  
(私たちや実現象)

- 実験の計画と実施
- 結果の評価と解釈
- 実験化学者との協働
- 理論化学者との協働



# 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！

「理解」や「発見」したいのは機械ではなく私たち人間  
つまり、自然法則の問題ではなく **私たち自身の精神と世界のあり方の問題**を問うことになる！

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

「理解」や「発見」したいのは機械ではなく私たち人間

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

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

「理解」や「発見」したいのは機械ではなく私たち人間  
つまり、自然法則の問題ではなく **私たち自身の精神と世界のあり方の問題**を問うことになる！

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

「理解」や「発見」したいのは機械ではなく私たち人間

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

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

「理解」や「発見」したいのは機械ではなく私たち人間

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

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

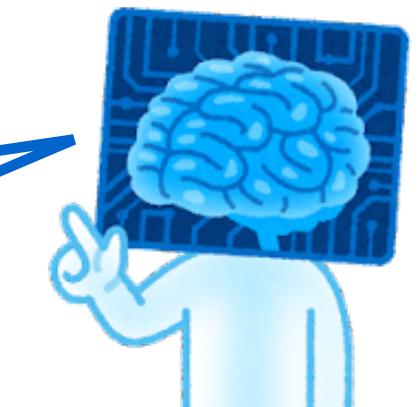
必要な情報のうち、**いつも偏った「一部」**しかデータにはできない前提で、私たち自身の許容限界に合う情報や示唆を得るために **「データを予測に変える道具」**をどう使えるか

「理解」や「発見」したいのは機械ではなく私たち人間

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

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

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

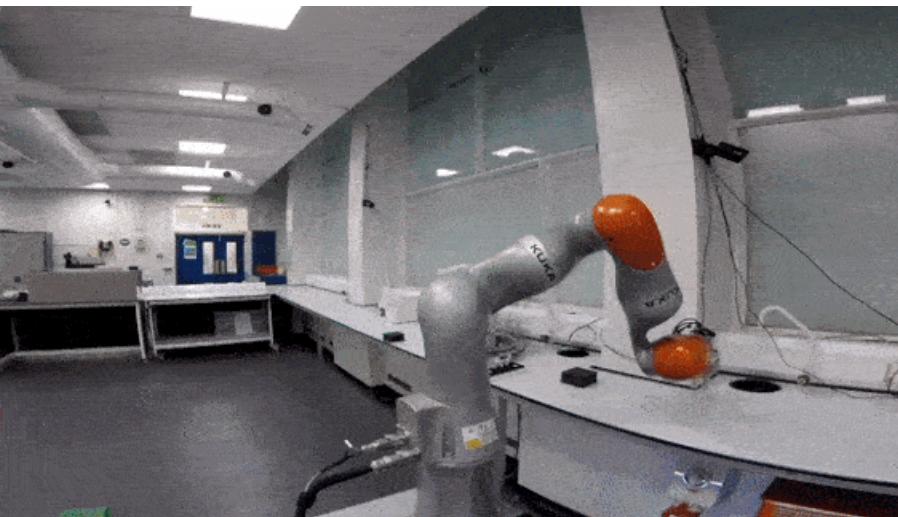


# 機械学習から機械発見へ

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



Science 363 (2019)



Nature 583 (2020)



Nature Reviews Drug Discovery 17 (2018)



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



Science 363 (2019)



Nature 583 (2020)



Nature Reviews Drug Discovery 17 (2018)



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

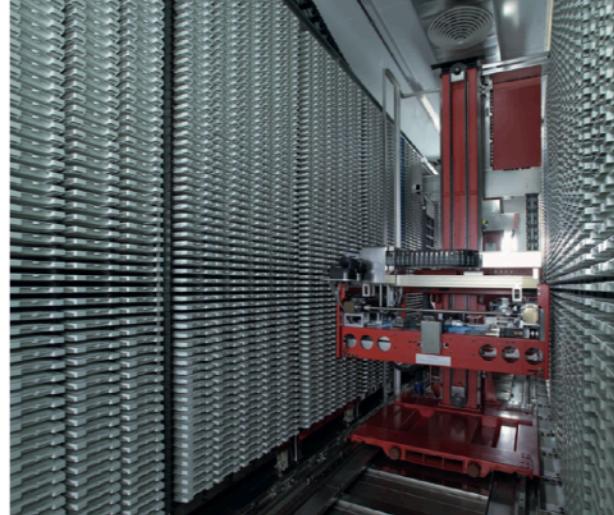
## 実験自動化の技術的発展：化学でも非効率な労働がいずれ自動化されるのは歴史的必然



Science 363 (2019)



Nature 583 (2020)



Nature Reviews Drug Discovery 17 (2018)



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

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



Science 363 (2019)



Nature 583 (2020)

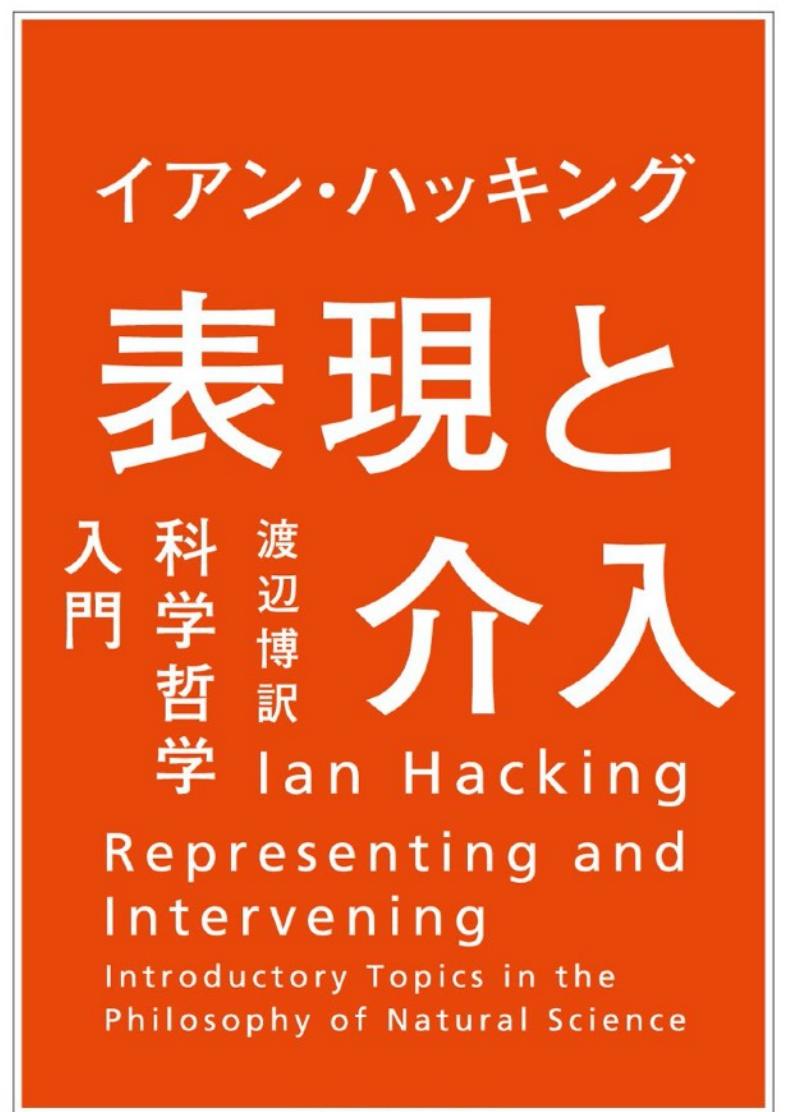
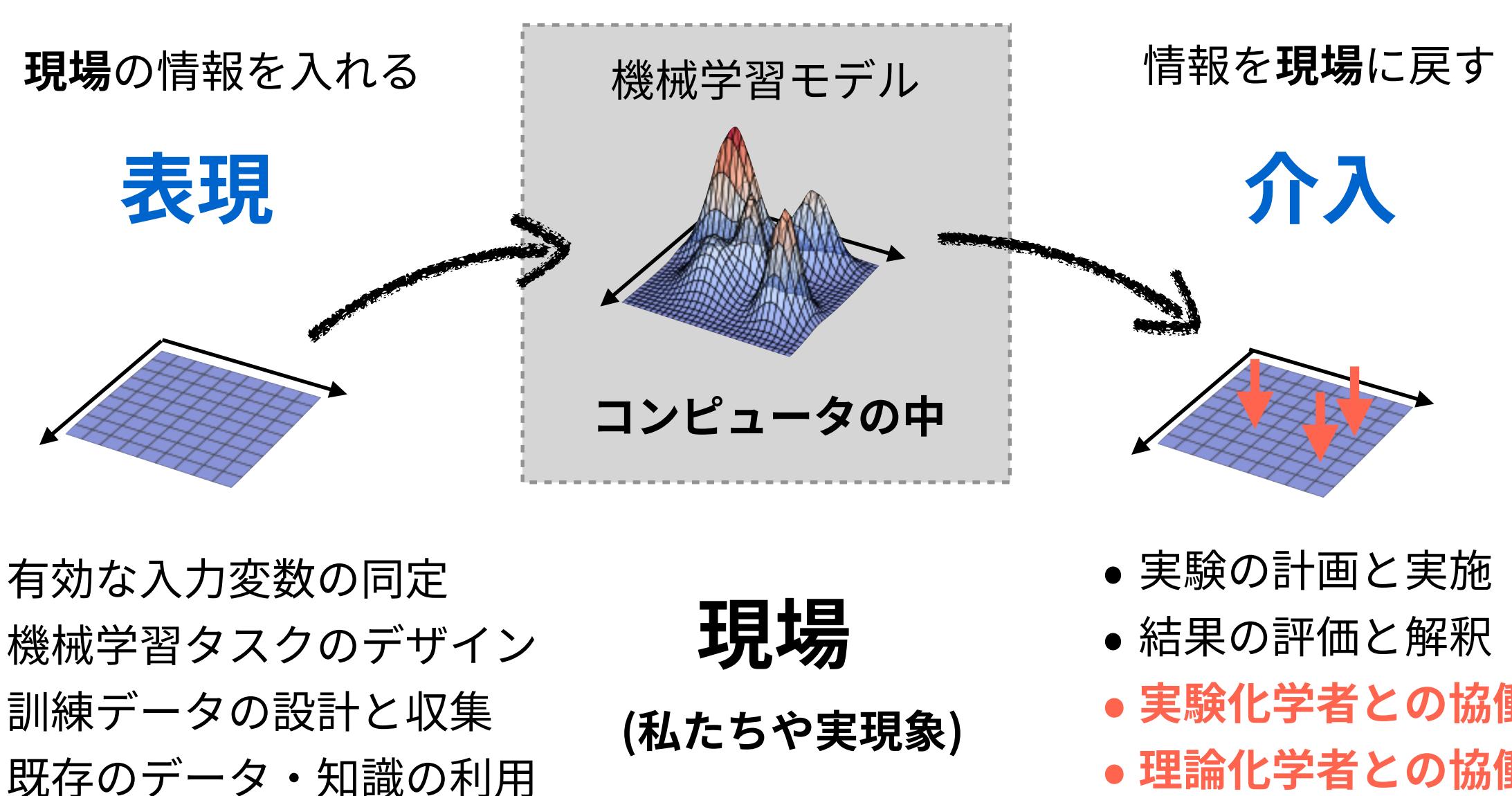


Nature Reviews Drug Discovery 17 (2018)



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

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



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

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



1. 機械学習は「データを予測に変える」
2. 機械学習は「新しい（そしてめっちゃ雑な！）コンピュータプログラムの作り方」

3. 現在の機械学習モデルはアホみたいにデータを食う…

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

4. 機械学習×化学の真の問題：機械学習から機械発見へ、予測から理解・発見へ  
シン
- 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

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」

