

不均一系触媒研究のための 機械学習と最適実験計画

理研CSRS インフォマティクス・データ科学推進プログラム成果報告会
2021年3月8日(月)：1:00pm - 5:30pm

たきがわ いちがく
瀧川 一学

<https://itakigawa.github.io/>

理化学研究所 革新知能統合研究センター(AIP) @京阪奈
北海道大学 化学反応創成研究拠点(WPI-ICReDD) @札幌
アイクレッド

自己紹介：瀧川一学 (たきがわ いちがく)

機械学習(ML)を研究している情報工学屋ですが X-informatics/
e-Scienceや色々な自然科学分野での利活用にも携わってきました！

自己紹介：瀧川一学 (たきがわ いちがく)

機械学習(ML)を研究している情報工学者ですが X-informatics/e-Scienceや色々な自然科学分野での利活用にも携わってきました！

10年 北大
(1995～2004)

情報科学の手法や理論の研究 (工学研究科)

Amazon(1994)
Google(1998)
2ちゃんねる(1999)
ヒトゲノム計画
完了(2003)

自己紹介：瀧川一学 (たきがわ いちがく)

機械学習(ML)を研究している情報工学者ですが X-informatics/e-Scienceや色々な自然科学分野での利活用にも携わってきました！

10年 北大
(1995~2004)

情報科学の手法や理論の研究 (工学研究科)

Amazon(1994)
Google(1998)
2ちゃんねる(1999)

7年 京大
(2005~2011)

バイオインフォマティクス (化学研究所)
ケモインフォマティクス (薬学研究科)

ヒトゲノム計画
完了(2003)
Facebook(2004)
YouTube(2005)
Twitter(2006)
iPhone(2007)
Android(2009)
Bitcoin(2009)

第4のパラダイム
(2009)
マテリアルゲノム
計画 (2011)

自己紹介：瀧川一学 (たきがわ いちがく)

機械学習(ML)を研究している情報工学者ですが **X-informatics/e-Science**や色々な自然科学分野での利活用にも携わってきました！

10年 北大
(1995~2004)

情報科学の手法や理論の研究 (**工学研究科**)

Amazon(1994)
Google(1998)
2ちゃんねる(1999)

7年 京大
(2005~2011)

バイオインフォマティクス (**化学研究所**)
ケモインフォマティクス (**薬学研究科**)

ヒトゲノム計画
完了(2003)
Facebook(2004)
YouTube(2005)
Twitter(2006)
iPhone(2007)
Android(2009)
Bitcoin(2009)

7年 北大
(2012~2018)

機械学習とデータ駆動科学 (**情報科学研究科**)
材料インフォマティクス (JSTさきがけ)

第4のパラダイム
(2009)

?年 理研/北大
(2019~)

細胞生物学 (理研AIP)
化学反応のデザインと設計 (北大ICReDD)

マテリアルゲノム
計画 (2011)
Society 5.0
(2016)

本日お伝えしたいこと

<https://itakigawa.github.io/news.html>

このスライドpdfはここにあります

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

本日お伝えしたいこと

<https://itakigawa.github.io/news.html>

このスライドpdfはここにあります

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

- MLがどういう技術なのか**MLの特性と限界**を正しく把握する

本日お伝えしたいこと

<https://itakigawa.github.io/news.html>

このスライドpdfはここにあります

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

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

本日お伝えしたいこと

<https://itakigawa.github.io/news.html>

このスライドpdfはここにあります

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

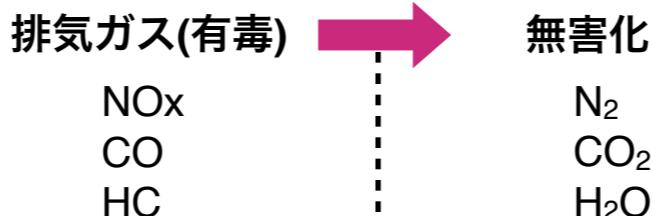
- MLがどういう技術なのか**MLの特性と限界**を正しく把握する
- 「データの収集計画(実験計画)と品質保証、適用範囲の理解」
が“**data-driven**”の心臓であることをいつも心に
- 「探索」が目的なら**MLの果たす役割はあくまで一部**と心得る
 - 👍 専門家との協働、分野の専門知識に照らした検証・解釈
 - 👍 シミュレーション・実験自動化・論理推論との融合

不均一系触媒 (Heterogeneous Catalysis)

触媒と反応物が固体と液体、固体と気体というように別の相
→ 「固体触媒表面上の気相反応」を本日は仮定します

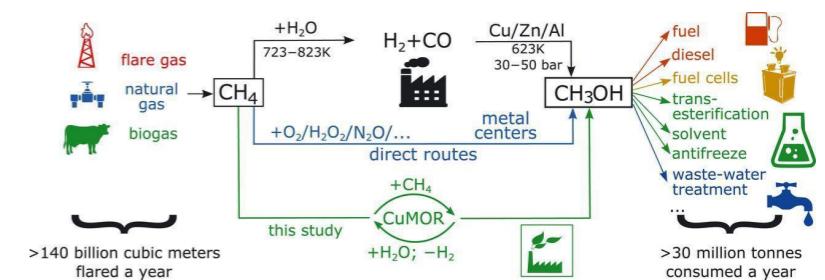
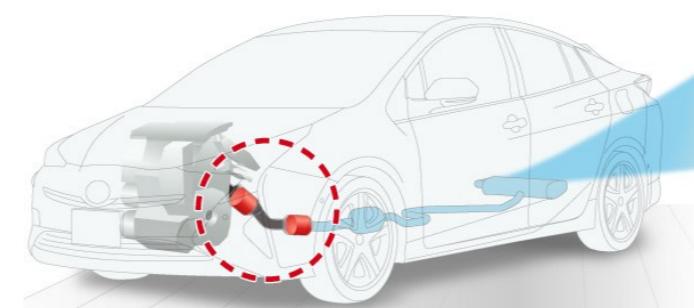
アンモニアの工業的合成 (ハーバー・ボッシュ法)

“水と石炭と空気からパンを作る方法”
20世紀の食糧難を解決した
人工的窒素固定



鉄系触媒など?

貴金属触媒など?
(Pt, Pd, Rh...)



- メタン \longrightarrow
- エタン
 - エチレン
 - メタノール
 - :

金属触媒 など?
(Li, 希土類, アルカリ土類)

不均一系触媒研究での機械学習の活用

JST CREST 革新材料開発 (細野領域)

触媒インフォマティクスの創成のための実験・理論・データ科学的研究



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



清水研一

高尾基史, 峯 真也, 鈴木慶介

高草木 達

鳥屋尾 隆

前野 禅

- Toyao+, *ACS Catalysis*. 2020. (Review)
- Liu+, *The Journal of Physical Chemistry C*. 2020.
- Suzuki+ *ChemCatChem*. 2019. (Front Cover) → この研究+その後を紹介
- Kamachi+ *The Journal of Physical Chemistry C*. 2019.
- Hinuma+ *The Journal of Physical Chemistry C*. 2018.
- Toyao+, *The Journal of Physical Chemistry C*. 2018
- Takigawa+ *RSC Advances*. 2016.

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



Review

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

pubs.acs.org/acscatalysis

Machine Learning for Catalysis Informatics: Recent Applications and Prospects

Takashi Toyao,^{†,‡} Zen Maeno,[†] Satoru Takakusagi,[†] Takashi Kamachi,^{‡,§} Ichigaku Takigawa,^{*,||,⊥} and Ken-ichi Shimizu^{*,†,‡}

Review Comments

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

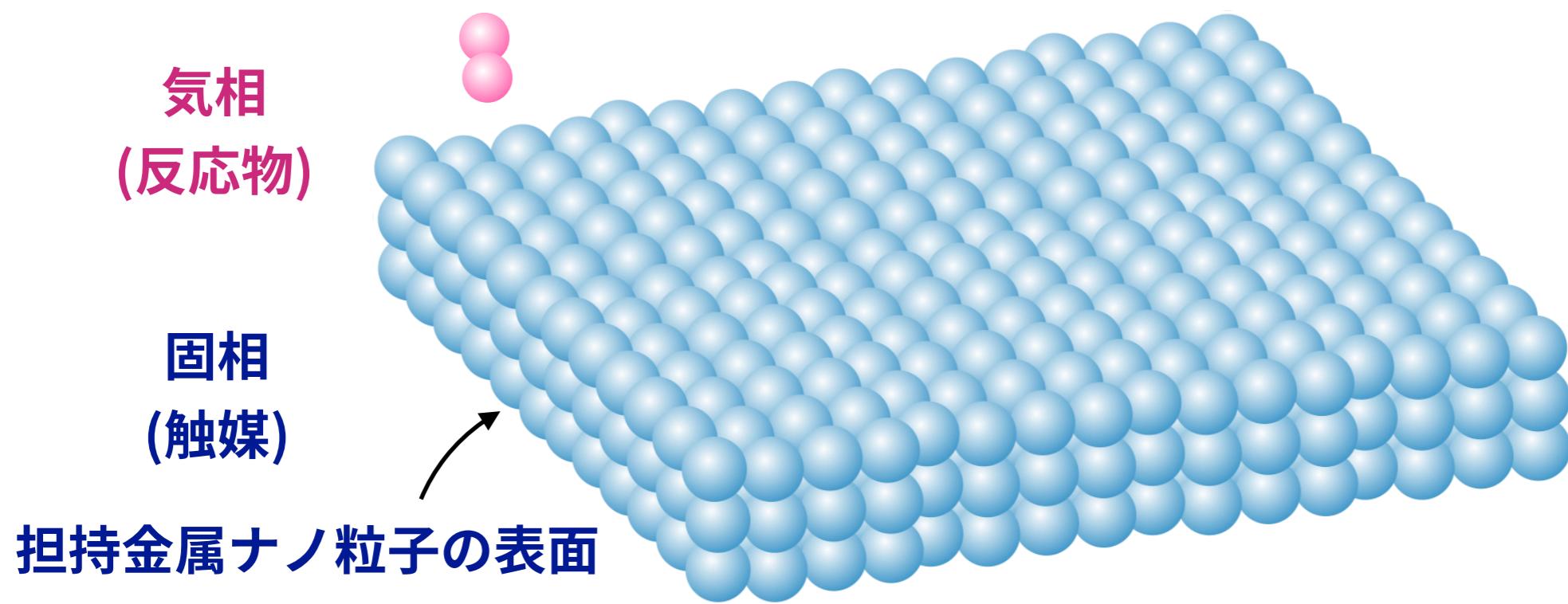
2章が機械学習のユーザガイド(数式なし)になっています！

本日の内容

1. はじめに：問題の難しさの確認
2. 機械学習をどう利活用するか
3. 文献から集めた実際の実験データ報告を使う
4. このとき機械学習に何が必要か
5. 実験計画と機械学習
6. おわりに：私が得た教訓

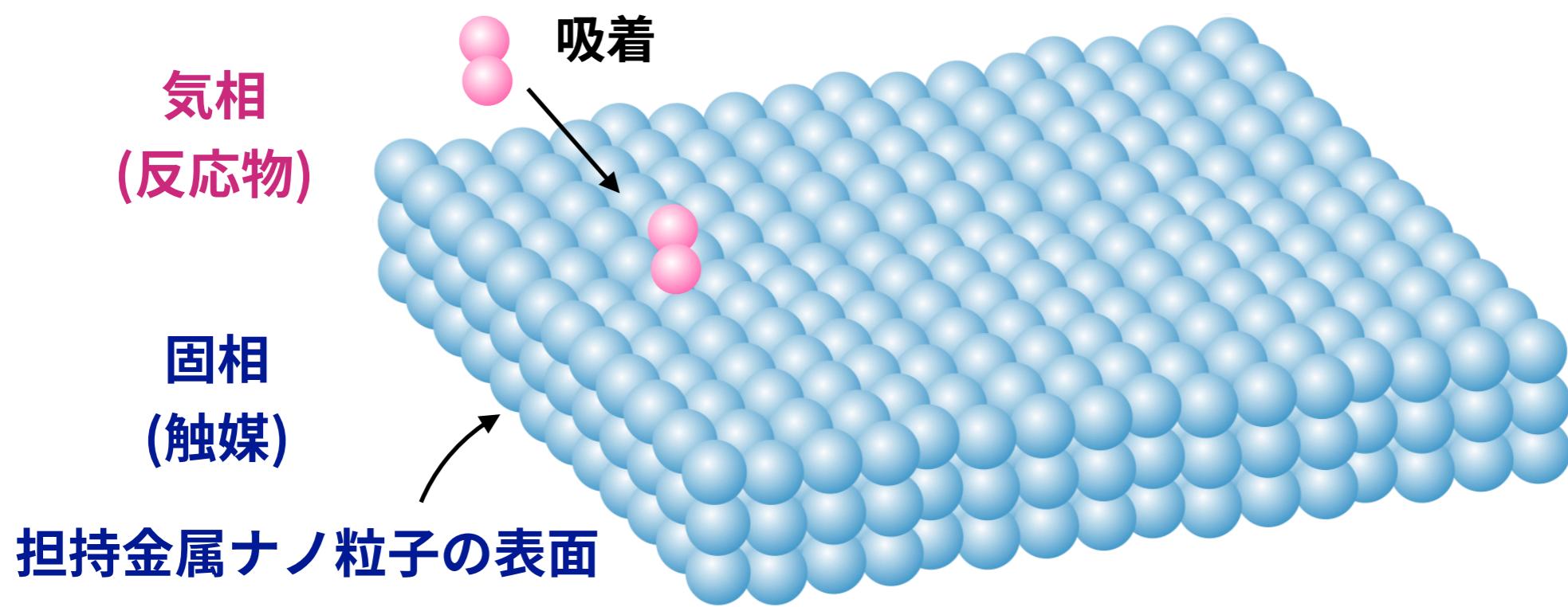
はじめに：問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



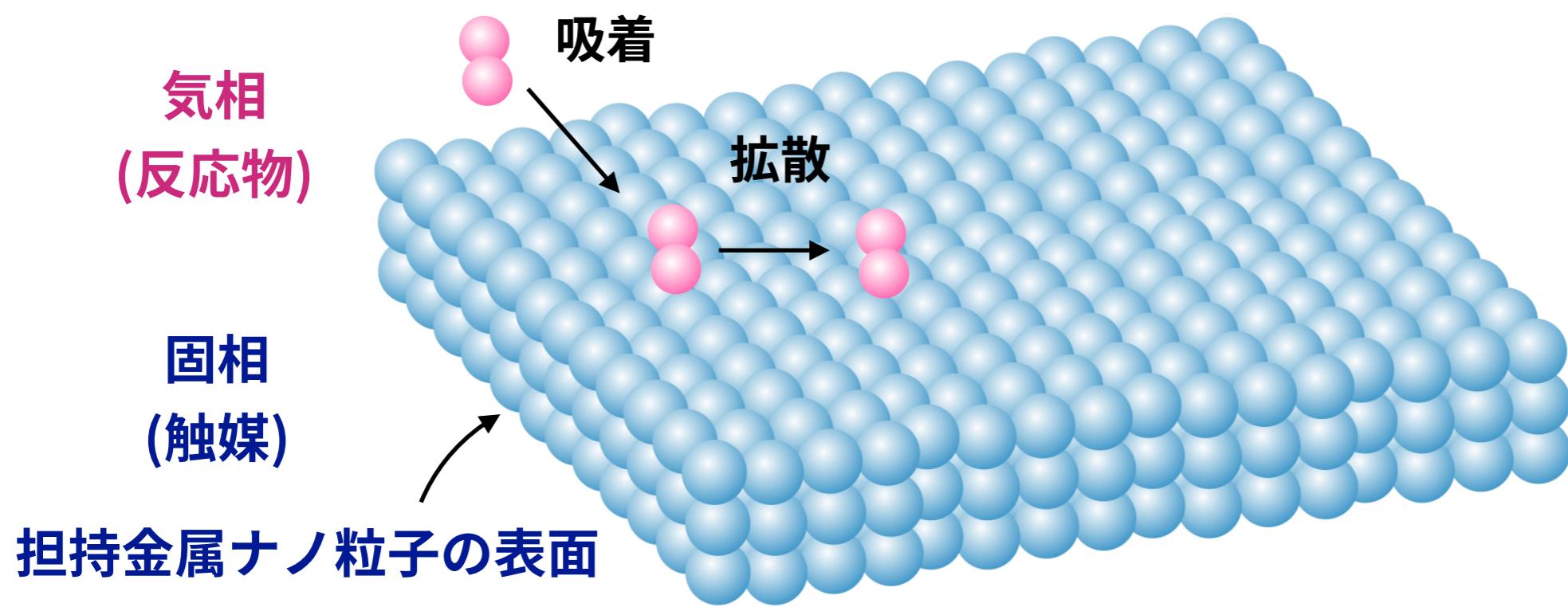
はじめに：問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



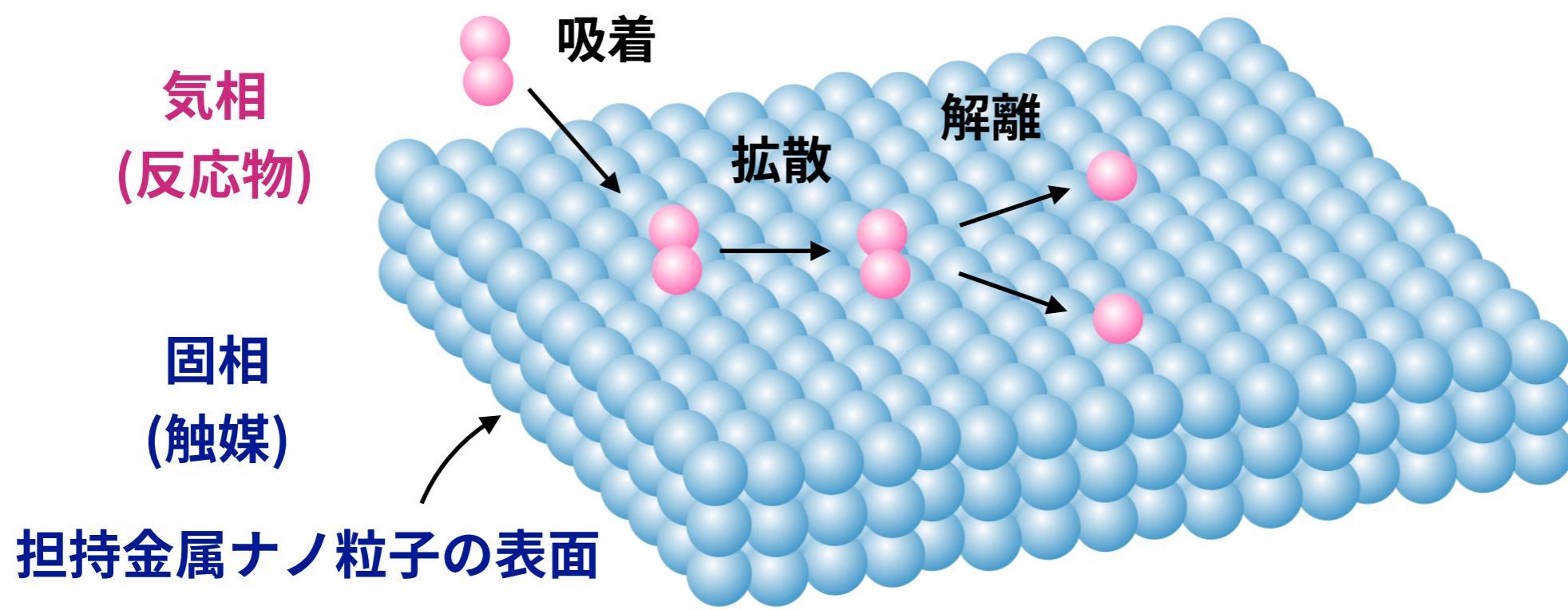
はじめに：問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



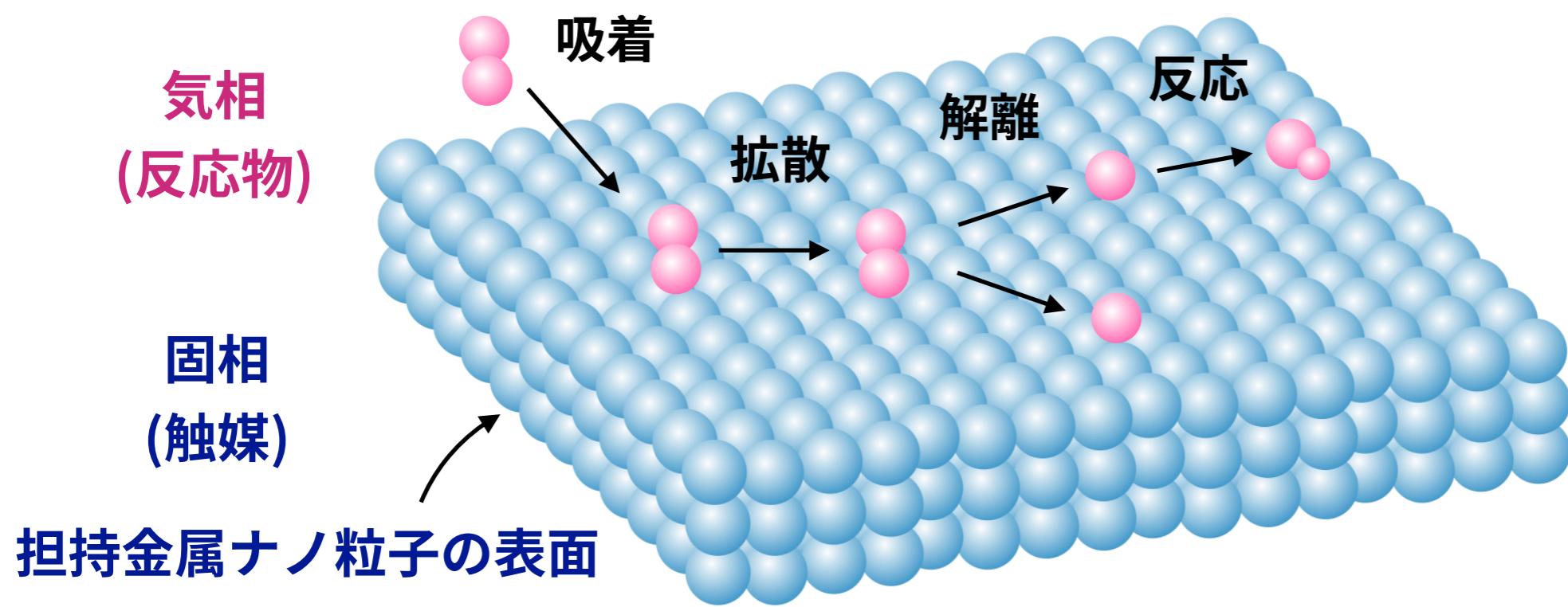
はじめに：問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



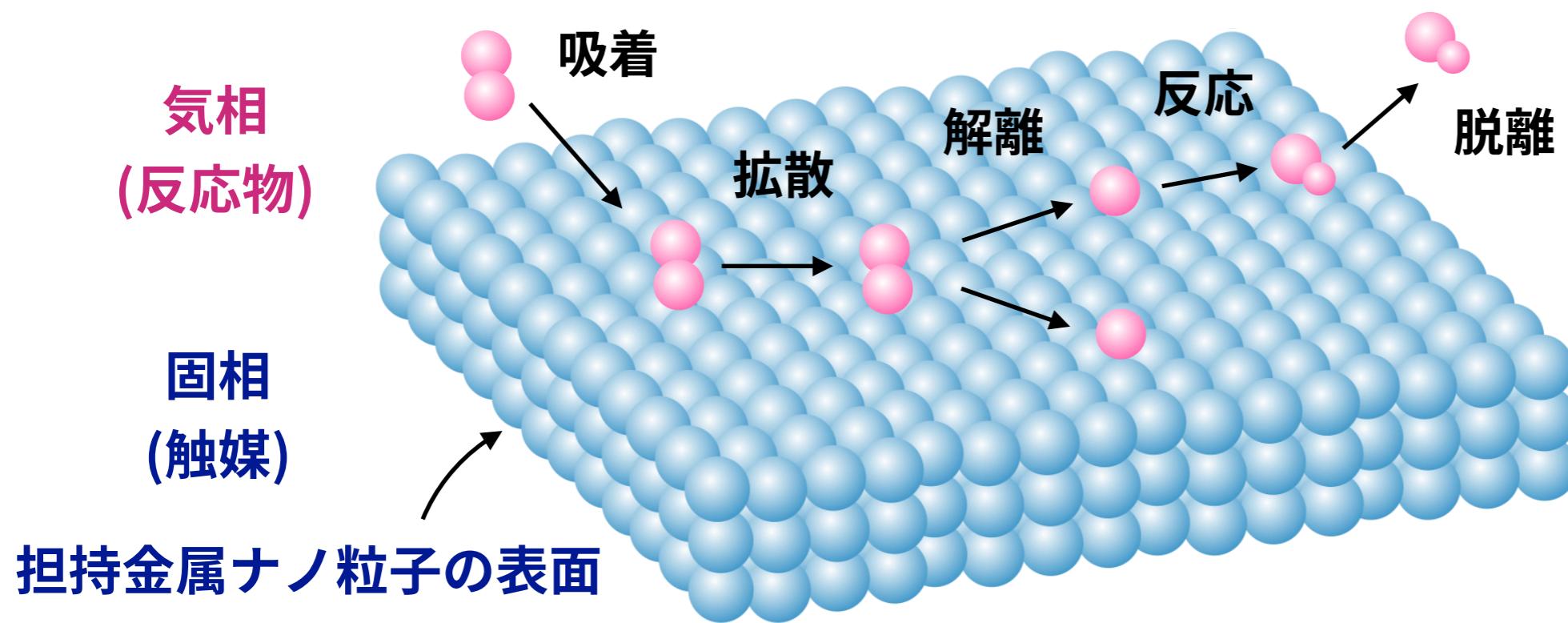
はじめに：問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



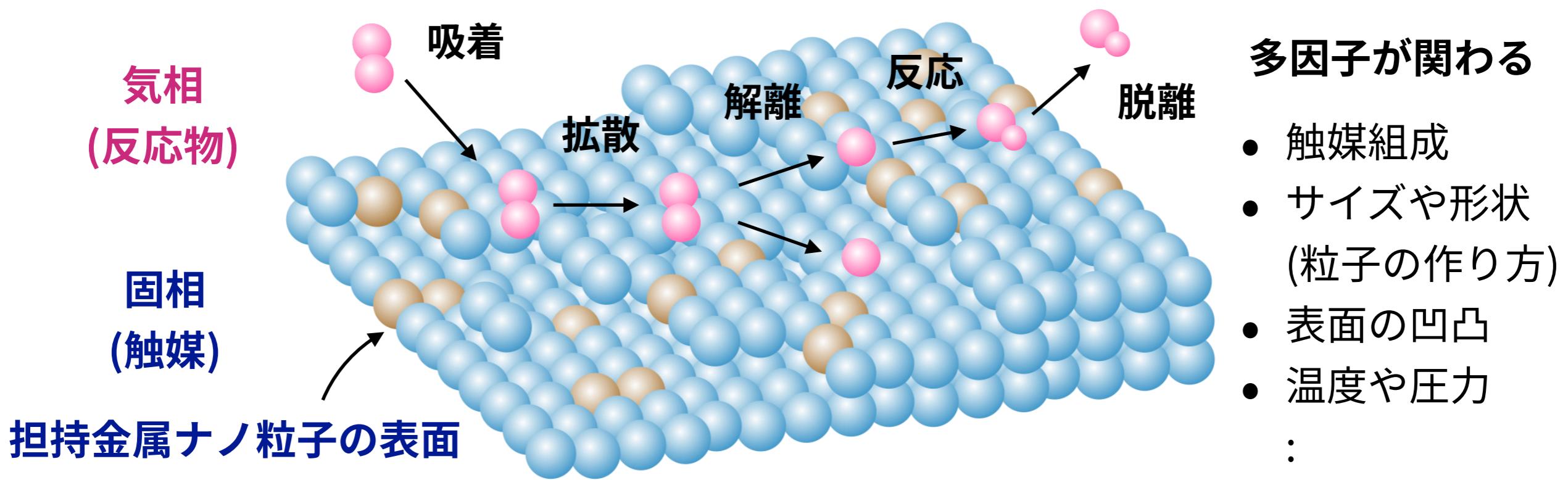
はじめに：問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



はじめに：問題の難しさの確認

「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ



はじめに：問題の難しさの確認

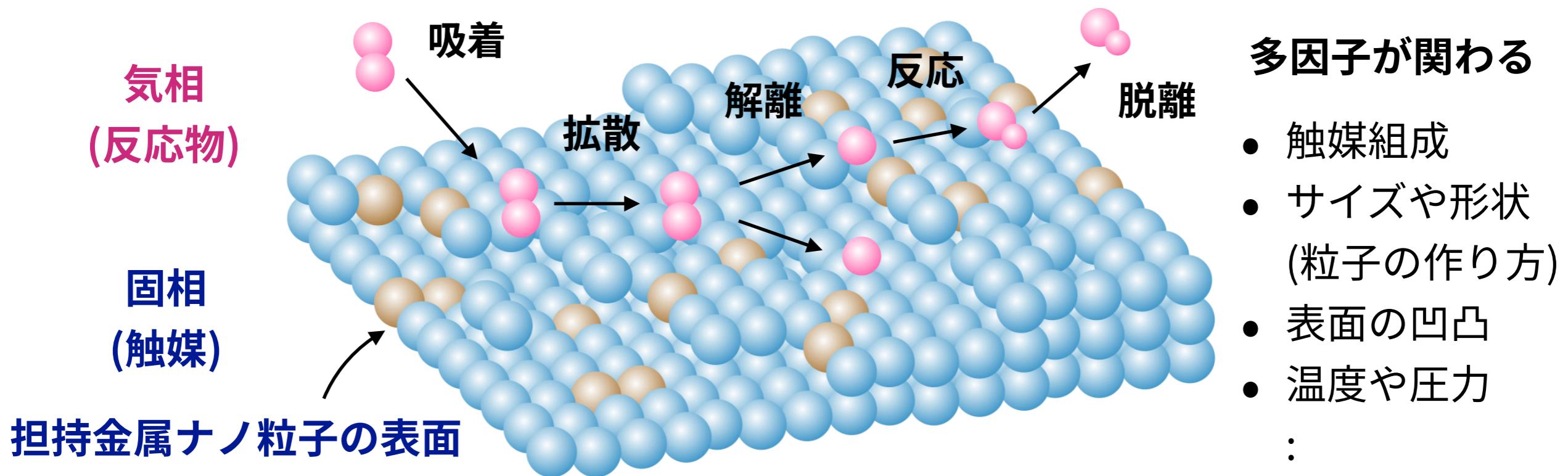
「固体触媒表面上の気相反応(複雑系)」の理解はそもそも激ムズ

→ そもそも「表面」が悪魔的な難しさ (“表面科学”)



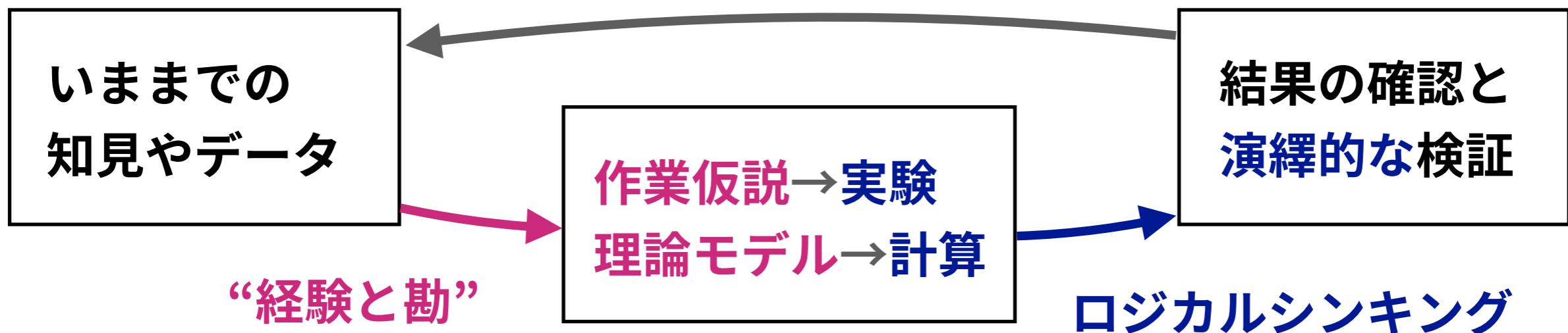
God made the bulk;
the **surface** was invented by the devil

パウリ大先生



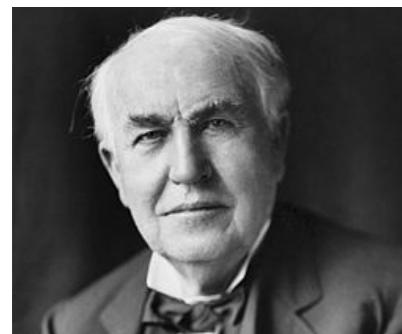
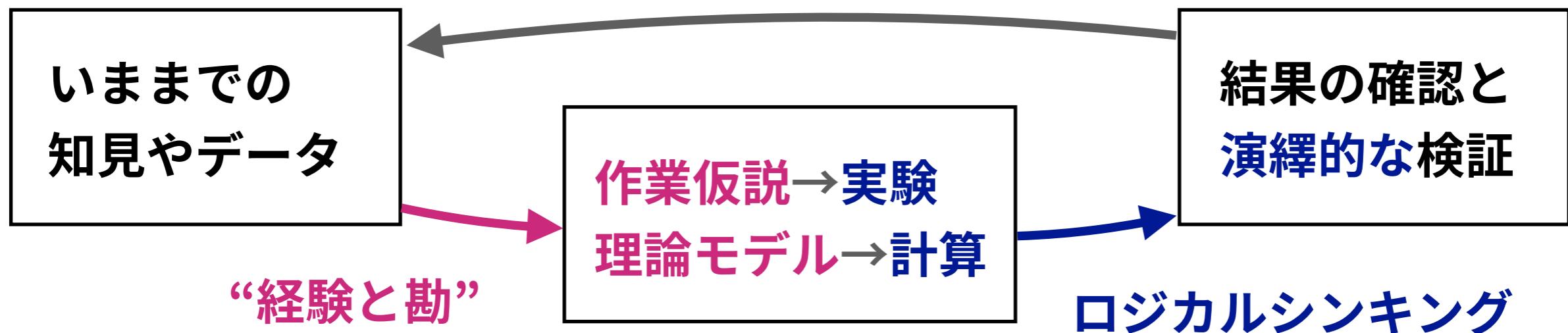
実験が主導して良い触媒が見つかってきたことは驚異的

仮説演繹法



実験が主導して良い触媒が見つかってきたことは驚異的

仮説演繹法 or “エジソン的な”経験論

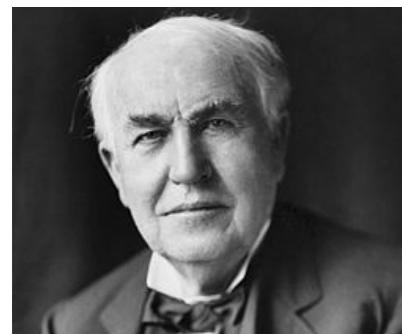
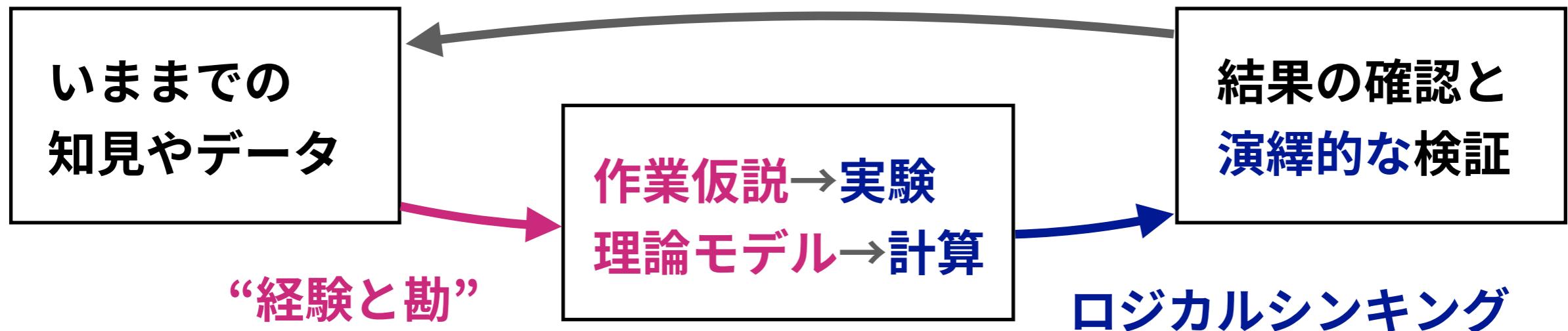


エジソン大先生

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

実験が主導して良い触媒が見つかってきたことは驚異的

仮説演繹法 or “エジソン的な”経験論



エジソン大先生

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

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

それでも！「発見」はものすごくレアイベントである

想定できる「触媒+実験条件+方法」の数は**天文学的に巨大**

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



“A needle in a haystack”



つまり「セレンディピティ ≠ 偶然の幸運」？

幸運は準備されたものにだけ降りる！！

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

つまり「セレンディピティ ≠ 偶然の幸運」？

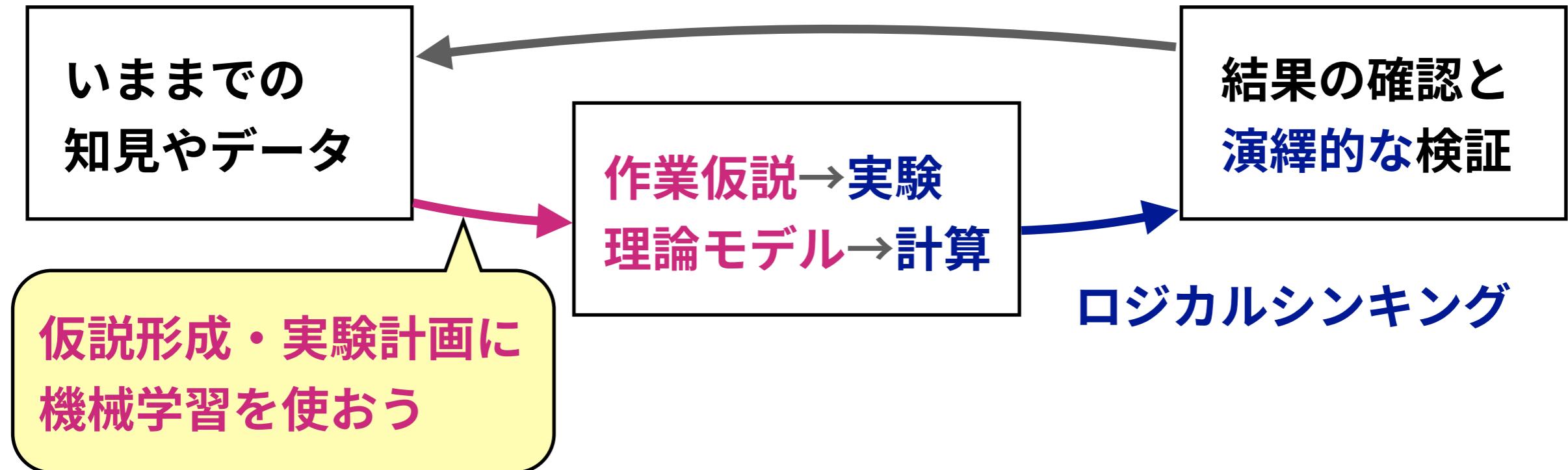
幸運は準備されたものにだけ降りる！！

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

➡ このあたりに“**data-driven**”が貢献できる余地がある

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

どういうデータで仮説形成・実験計画をdriveできるか？



- 優れた実験科学者に今までの全人生で入力された情報は膨大（データ化されない情報がほとんど）
- データ駆動：手に入るデータから近似的に迫るしかないが、その代わりに人間の認知キャパシティの制約から**自由になる**
⇄ 人間は多数の因子の複雑な多次元相関を把握できない

機械学習をどう活用するか

機械学習を活用するためには

機械学習モデルを訓練するための「データ」をどうするかが鍵

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

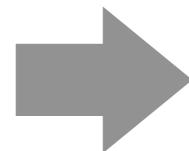
文献から集めた実際の実験データ報告を使う

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

現実をすべてモデル化することは不可能

→ その定義からして“モデル”とは何らかの捨象や近似を含む。

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



文献から集めた実際の実験データ報告を使う

「過去に報告された現実を見てみよう」

文献から集めた実際の実験データ報告を使う

対象はメタンの酸化カップリング反応、目的変数はC₂収率

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



Nr of public	A	B	C	N	O	R	S	T	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM
	Cation 1	Cation 1	Anion 1	Anion 1	Promotor	Support	Support	Preparati	Temperat	p(CH ₄), bar	p(O ₂), bar	p(CH ₄)/p(O ₂)	p total,	Contact time, s	X(O ₂), %	X(CH ₄), %	S(CO _x), %	S(C ₂ ⁻), %	S(C ₂ ⁻), %	S(C ₂), %	Y(C ₂), %		
1	Mn	9.2				AI	90.8	Impregnat	1073	0.40	0.08	4.8	1.0		0.04	11.0				45.5	5.0		
20	Li	30.3				n.a.		993	0.08	0.04	2.0	1.0		5.30	85.0	38.0				50.0	19.0		
21	Mg	66.7	S	33.3				Impregnat	1019	0.65	0.08	8.1	1.0		1.40	39.0	4.0		23.0	41.0	64.0	2.6	
22	Mg	55.0	S	45.0				Impregnat	1017	0.66	0.08	8.3	1.0		3.00	65.0	10.0		27.0	40.0	67.0	6.7	
23	Na	7.0	S	60.0				Impregnat	1017	0.64	0.08	8.0	1.0		0.19	39.0	3.0		23.0	19.0	42.0	1.3	
75	Pb	20.0				AI	80.0	n.a.	1030	0.96	0.05	19.2	1.0		0.40	100.0	6.8		17.6	32.8	50.4	3.4	
76	Pb	20.0				Si	80.0	n.a.	1103	0.96	0.05	19.2	1.0		0.55	44.1	18.7		18.7	20.5	39.2	7.3	
486	K	3.0	Cl	3.0	Cl	AI	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		30.8				33.6	10.3	
487	Li	3.0	Cl	3.0	Cl	AI	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		2.2				76.9	1.7	
488	Ba	3.0	Cl	6.0	Cl	AI	82.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		32.1				32.1	10.3	
489	Na	3.0	Cl	3.0	Cl	AI	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		35.0				33.5	11.7	
490	Cs	3.0	Cl	3.0	Cl	AI	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		30.2				24.3	7.3	
491	Ag	18.0			Cl	AI	82.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		20.4				0.0	0.0	
492	Ag	18.0	C	41.0	Cl			Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.8				0.3	0.1	
493	Pr	5.0			Cl	AI	86.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.6				0.2	0.1	
494	Pr	1.0			Cl	AI	90.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.1				0.0	0.0	
495	Bi	1.0			Cl	AI	81.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		23.4				1.3	0.3	
496	Ba	1.0			Cl	AI	81.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		27.8				0.7	0.2	
497	Ba	5.0			Cl	AI	77.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.0				1.7	0.4	
498	K	3.0			Cl	AI	79.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		17.2				23.3	4.0	
499	Ba	3.0	Cl	6.0	Cl	AI	82.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		15.8				28.0	4.4	
500	Ba	3.0	Cl	6.0	Cl	AI	73.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		27.1				30.4	8.2	
501	Ca	1.0	Cl	2.0	Cl	AI	79.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		15.8				25.4	4.0	
502	Ag	18.0			Cl	AI	82.0	Therm.dec	973	0.10	0.05	2.0	1.0		1.50		5.0				0.0	0.0	
503	Ba	3.0	Cl	6.0	Cl	AI	73.0	Therm.dec	973	0.10	0.05	2.0	1.0		1.50		17.2				25.4	4.4	
504	Ba	3.0	Cl	6.0	Cl			Therm.dec	973	0.10	0.05	2.0	1.0		1.50		26.7				15.3	4.1	
505	Ba	3.0	Cl	6.0	Cl	AI	73.0	Therm.dec	973	0.10	0.05	2.0	1.0		1.50		21.3				30.4	6.5	
506	Sr	3.0	Cl	6.0	Cl	AI	91.0	Impregnat	1023	0.10	0.05	2.0	1.0		1.50		30.3				56.0	17.0	
507	Ba	28.0	C	28.0	Cl	AI	44.0	Impregnat	1023	0.10	0.05	2.0	1.0		1.50		43.2				41.8	18.1	
508	Ba	2.0	Cl	2.0	Cl	AI	21.0	Impregnat	1010	0.00	0.05	1.0	1.0		0.10		17.0				11.0	10.0	

このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力 x

- 元素組成比
- 実験条件

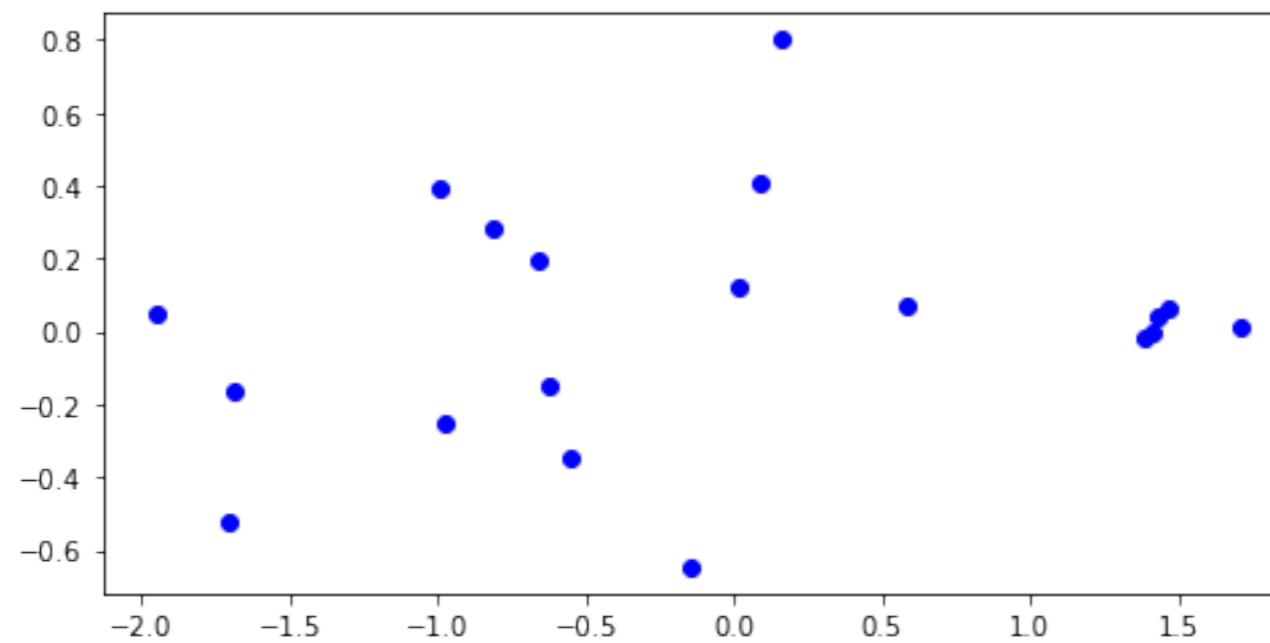
機械学習モデル

出力 y

- 収率

訓練データ

出力 y



入力 x

このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力 x

- 元素組成比
- 実験条件

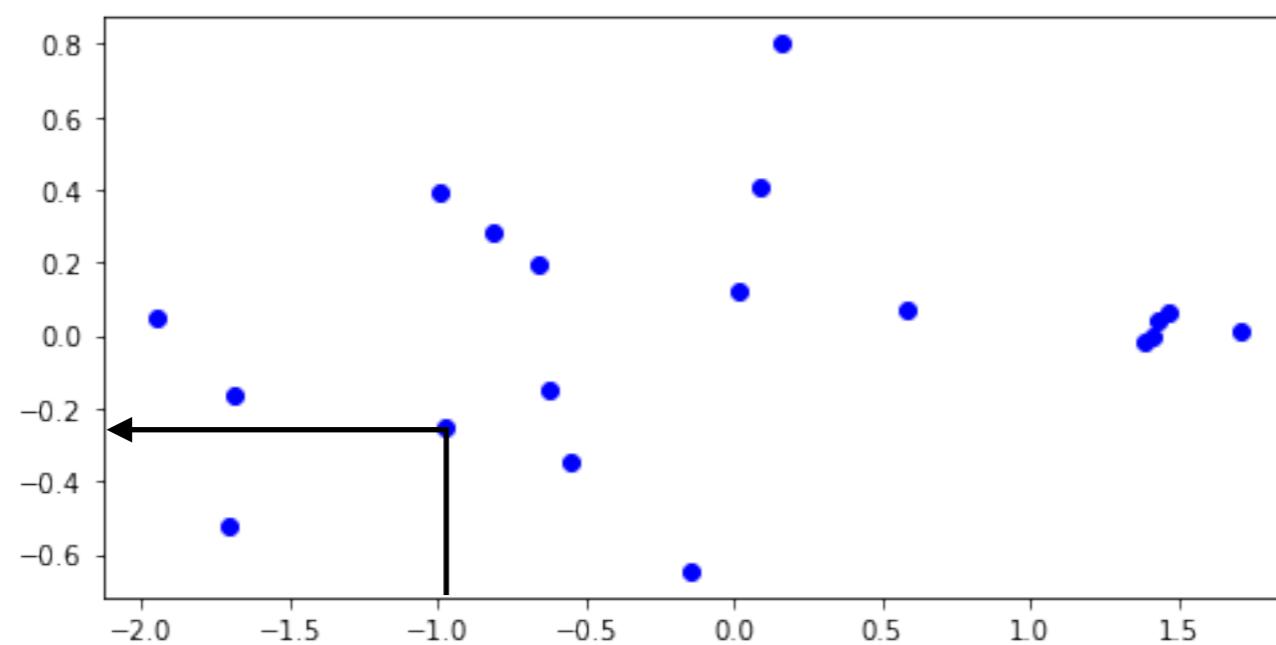
機械学習モデル

出力 y

- 収率

訓練データ

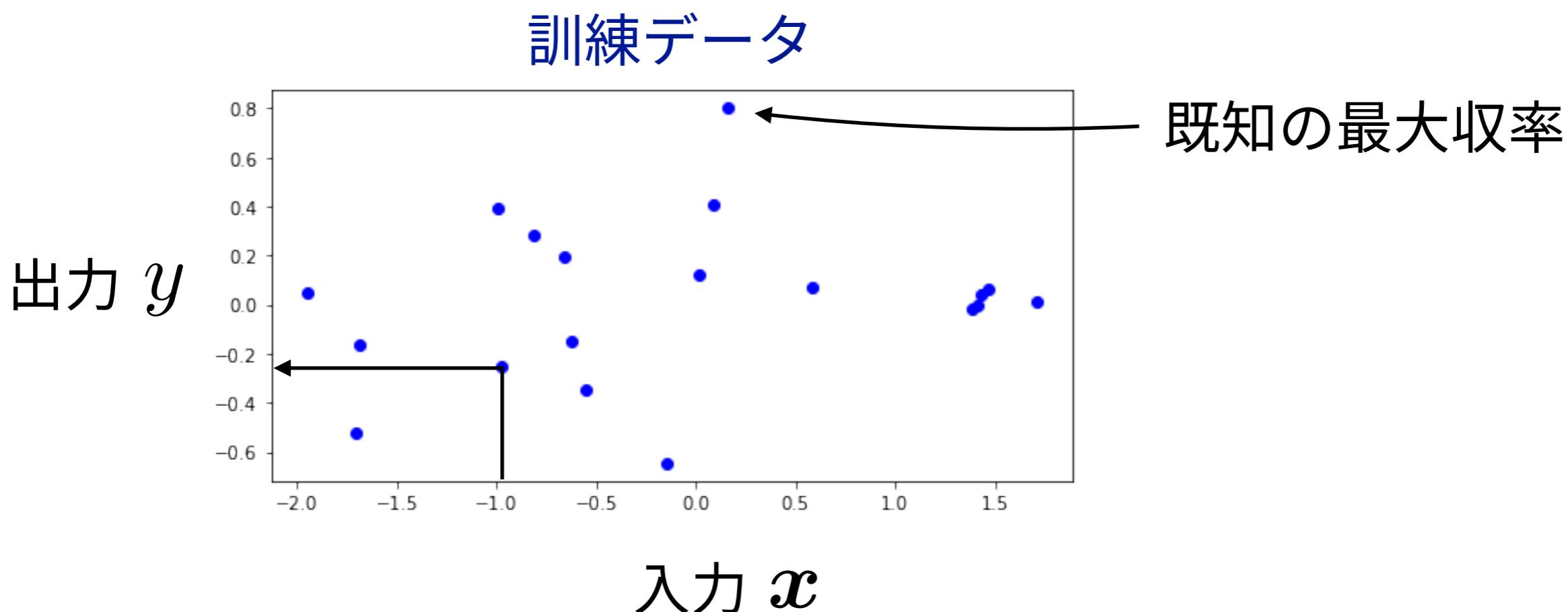
出力 y



入力 x

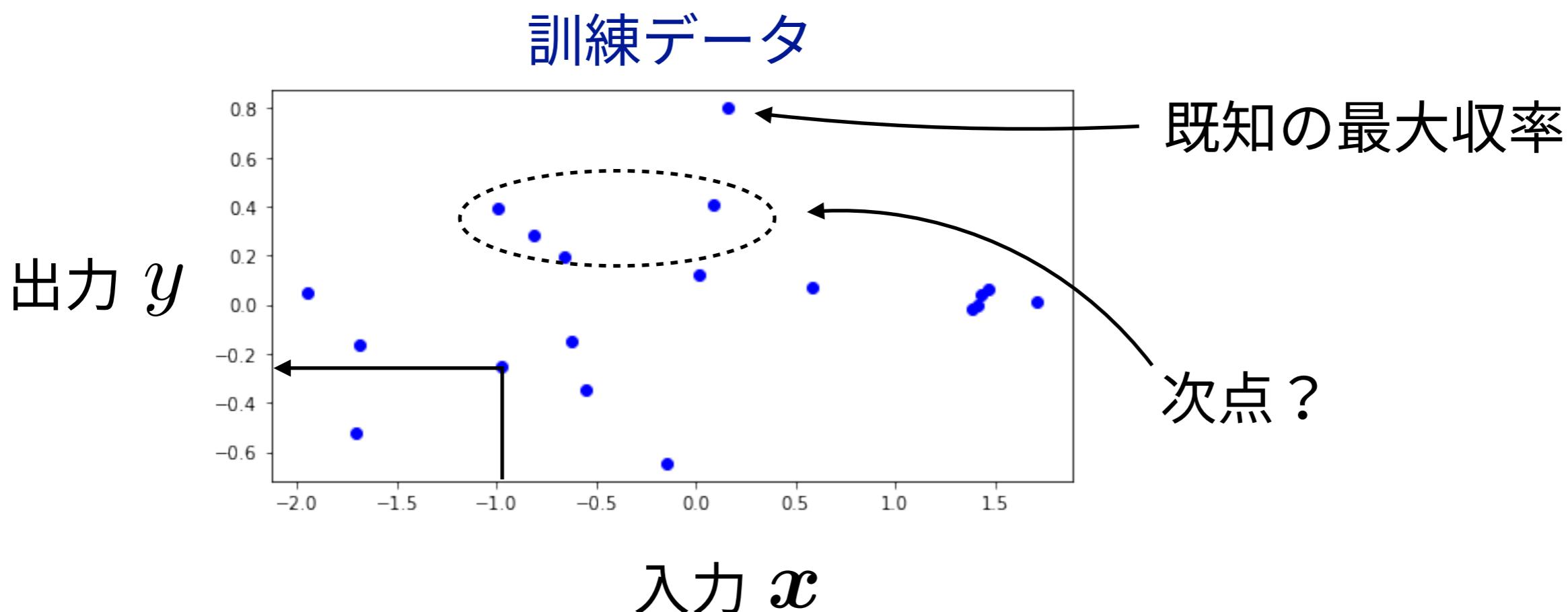
このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



このとき機械学習に何が必要か

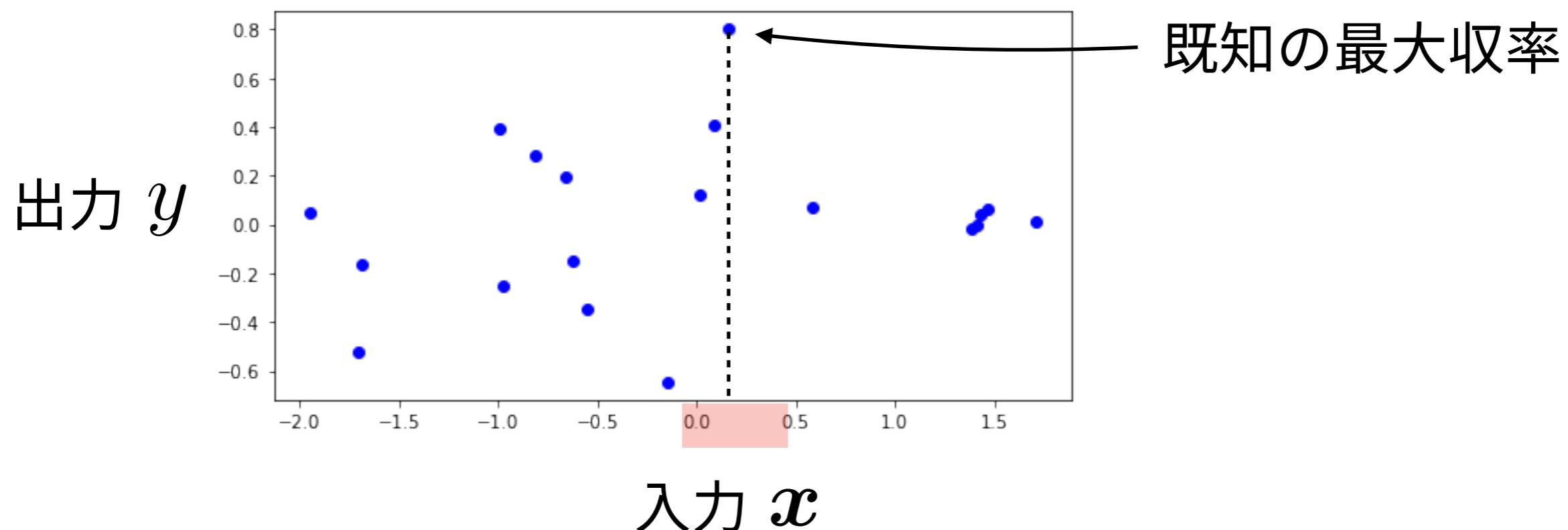
この表データで機械学習モデルを訓練し予測すれば良い？



知りたいことに応じて機械学習の方法や注意点が変わる

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

目的=探索 (既知の触媒より良い触媒を見つけたい！)



このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力 x

- 元素組成比
- 実験条件

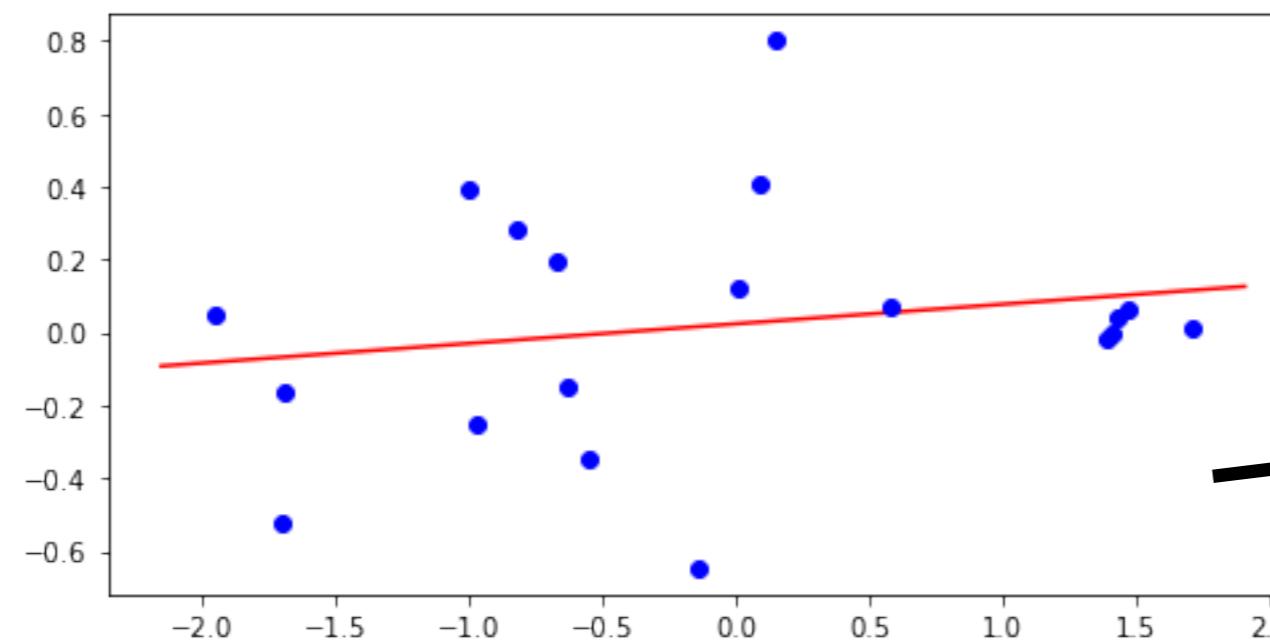
機械学習モデル

出力 y

- 収率

LinearRegression()

出力 y



入力 x

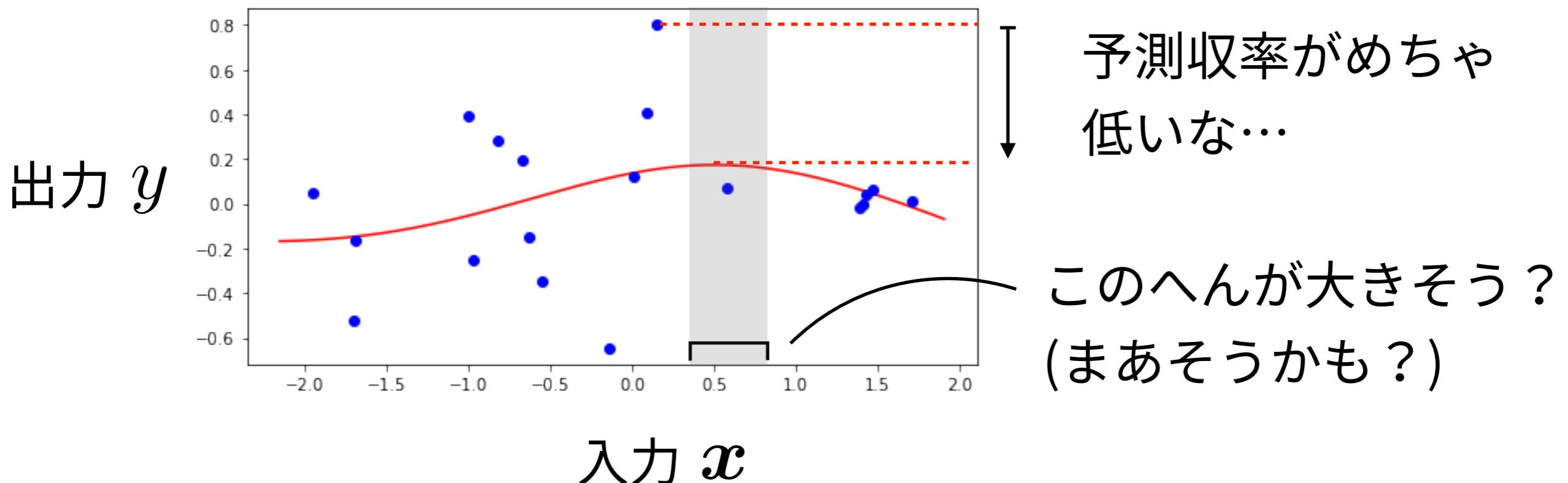
x を大きくすれば
良いという謎の
示唆しかくれない
(Underfit)

このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



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

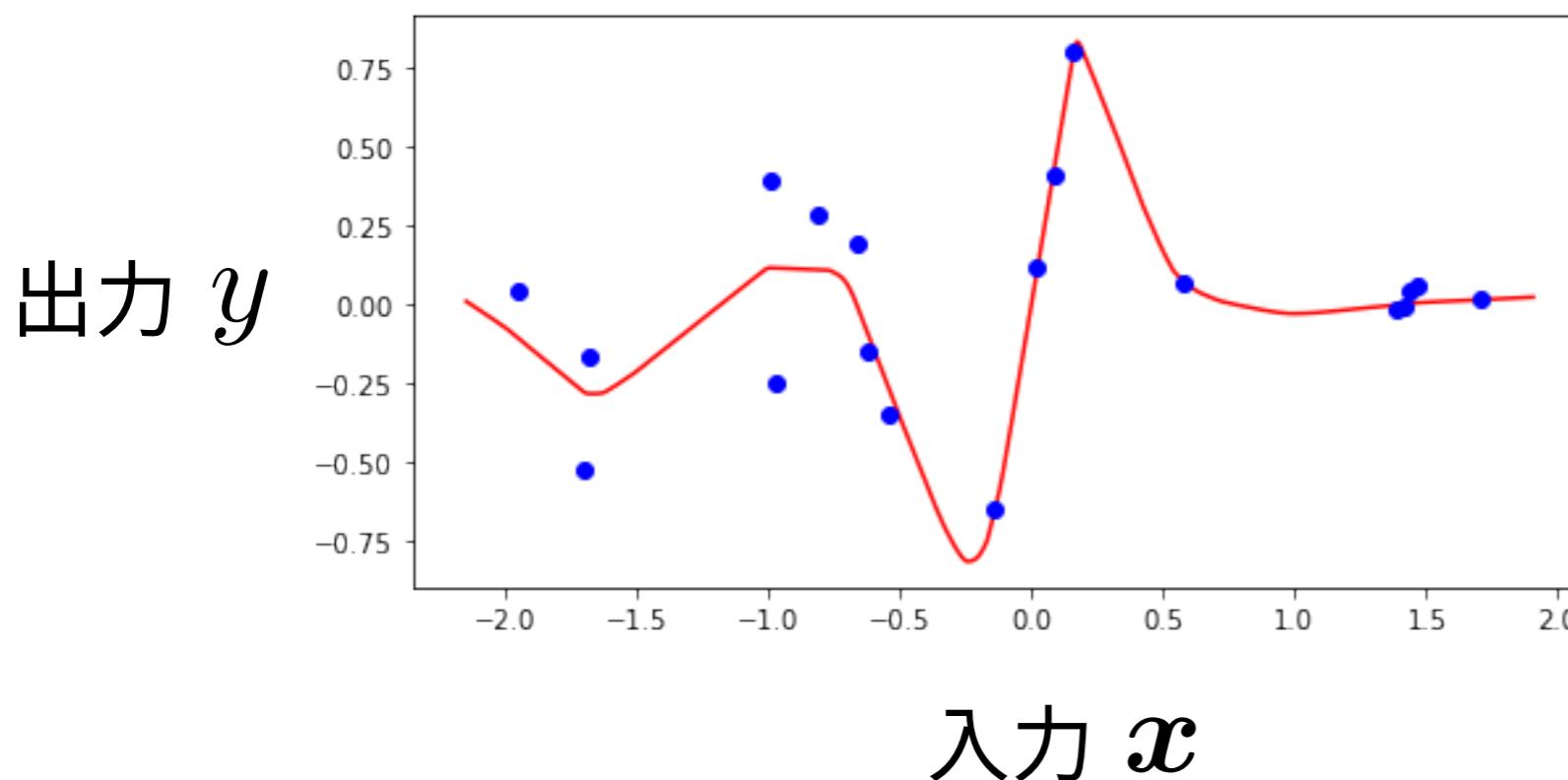


このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



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



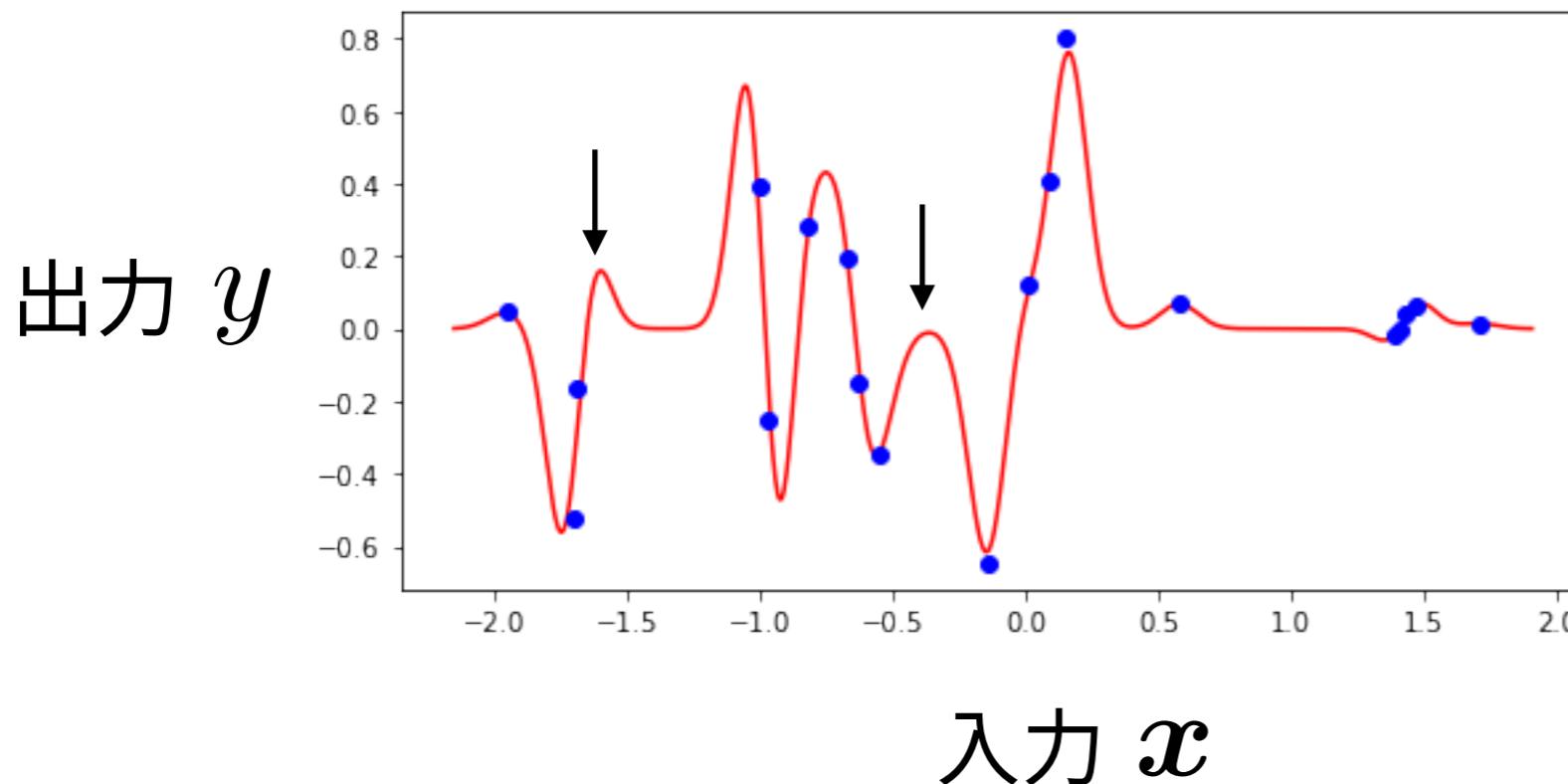
activationを「ReLU」に
(負値をゼロ置換)
うーん?
わりとよさそうかも?

このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



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



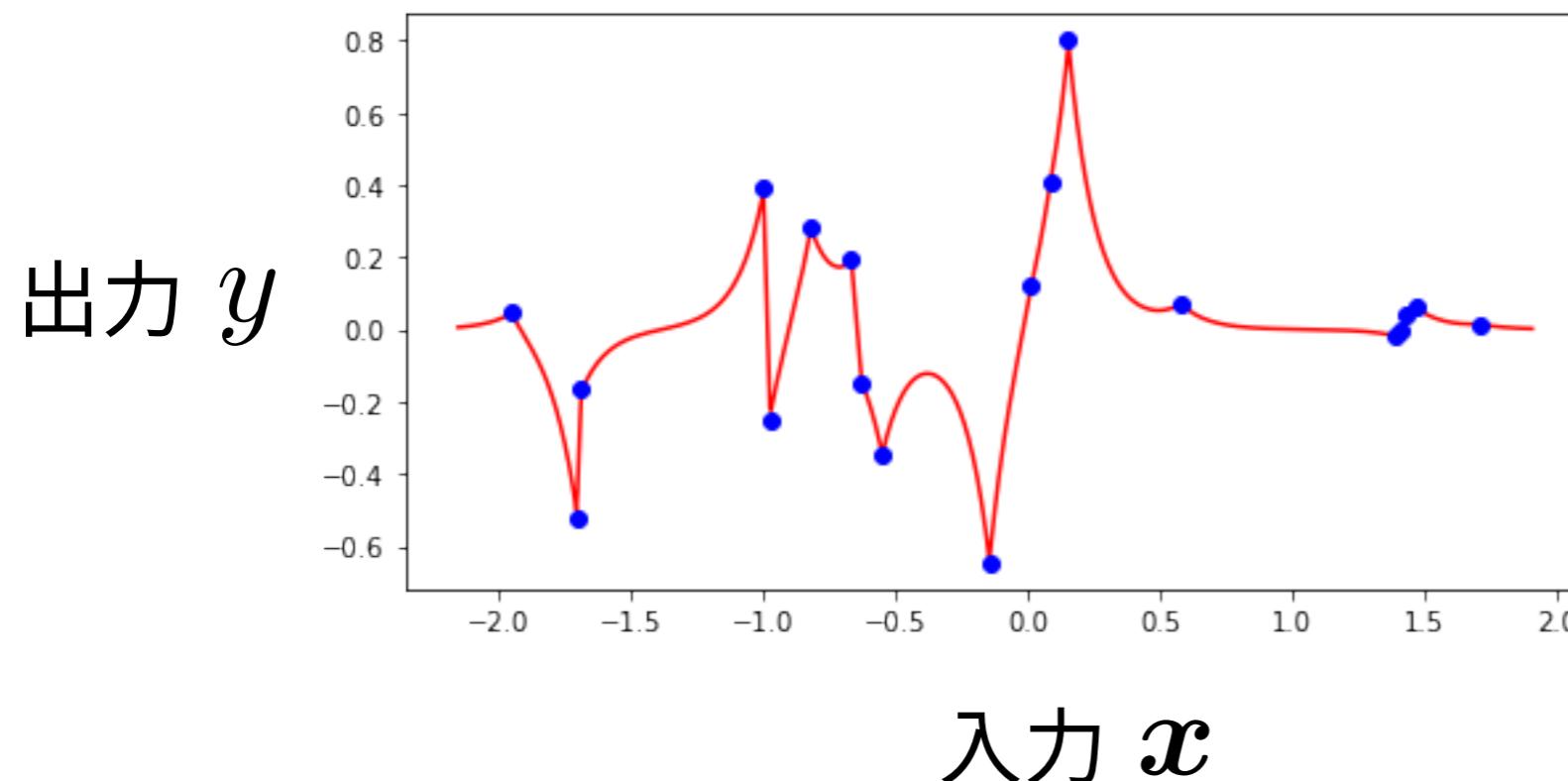
うーん？？
特に ↓ の箇所は
これで良いのか..！？
アーティファクト？

このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



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



うーん？？

このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力 x

- 元素組成比
- 実験条件

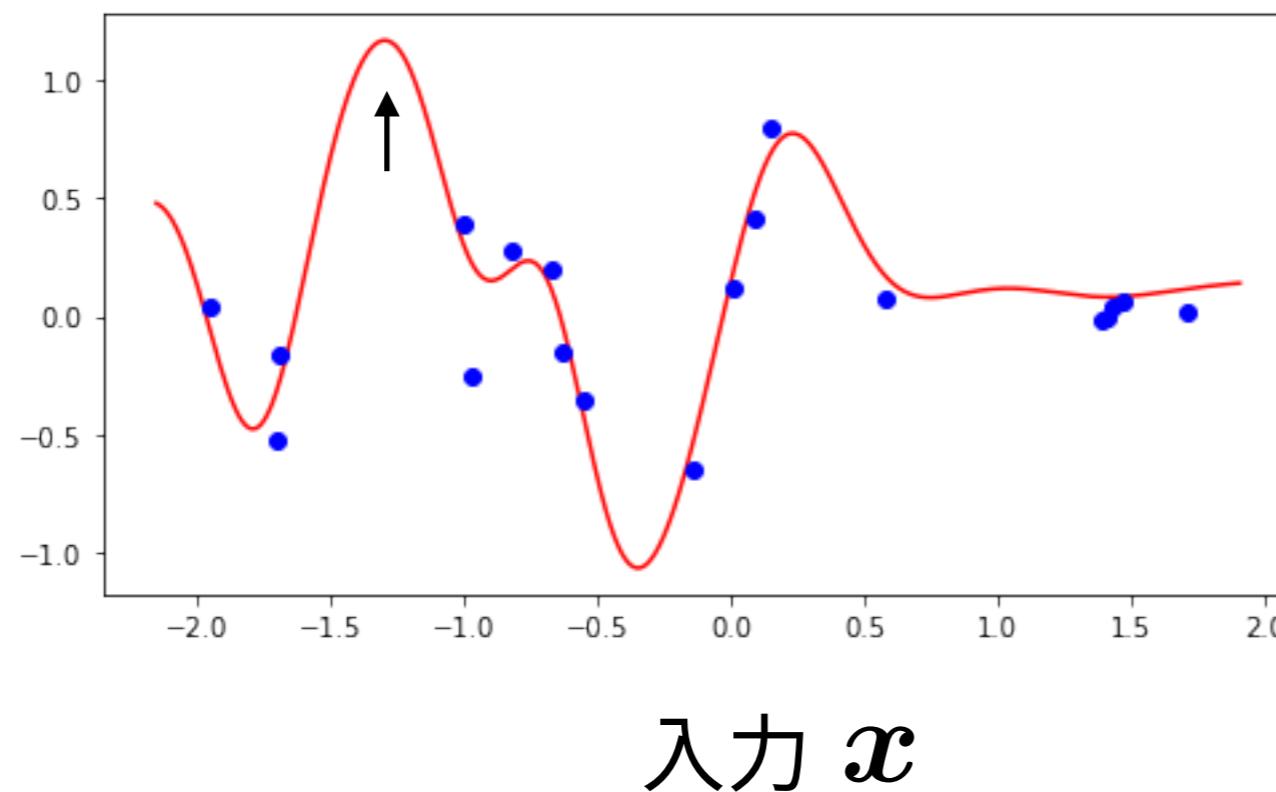
機械学習モデル

出力 y

- 収率

SVR(kernel='rbf', gamma=10, C=20)

出力 y



うーん？？？

特に ↑ の箇所は
これで良いのか..！？

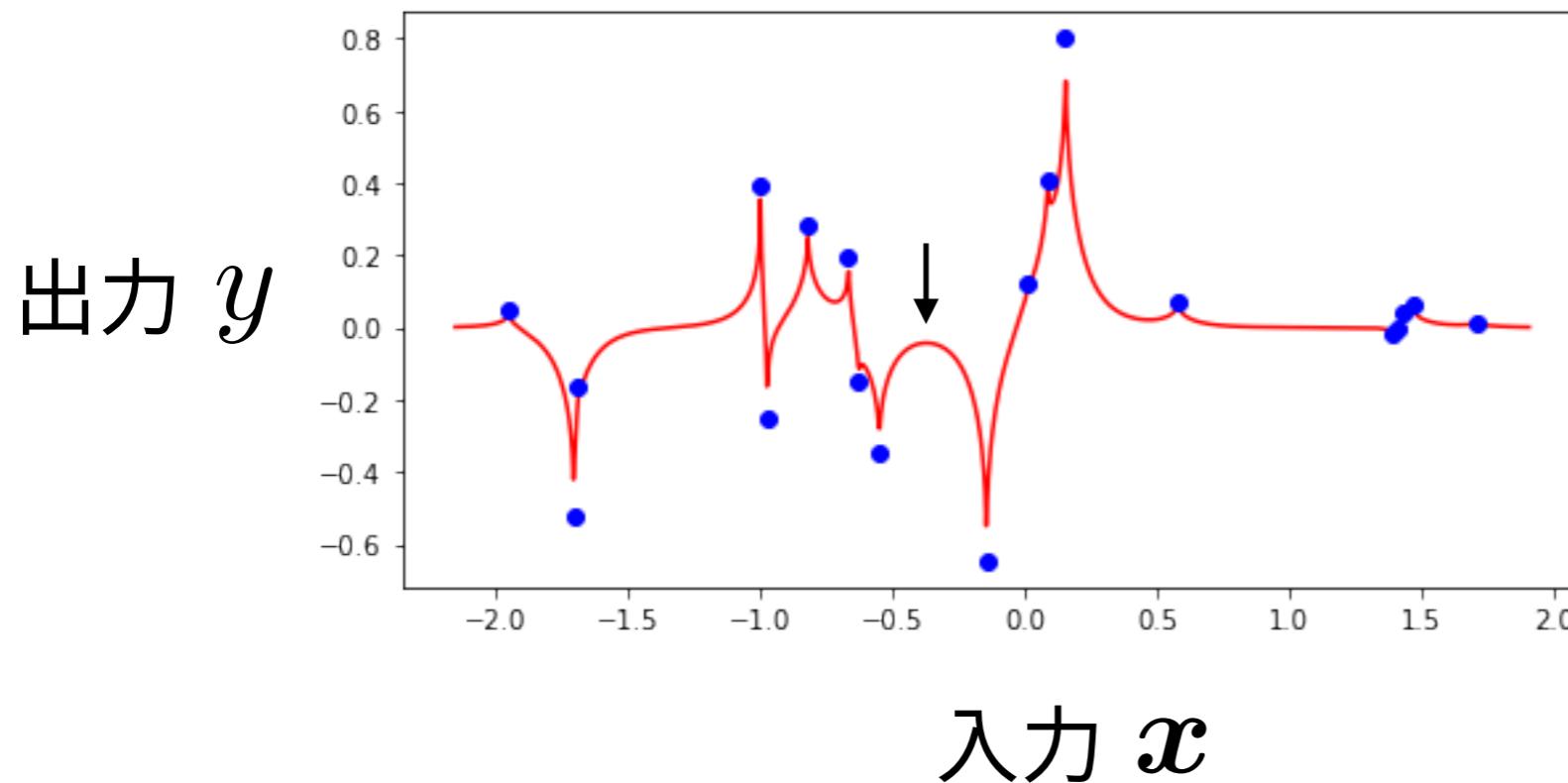
アーティファクト？

このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



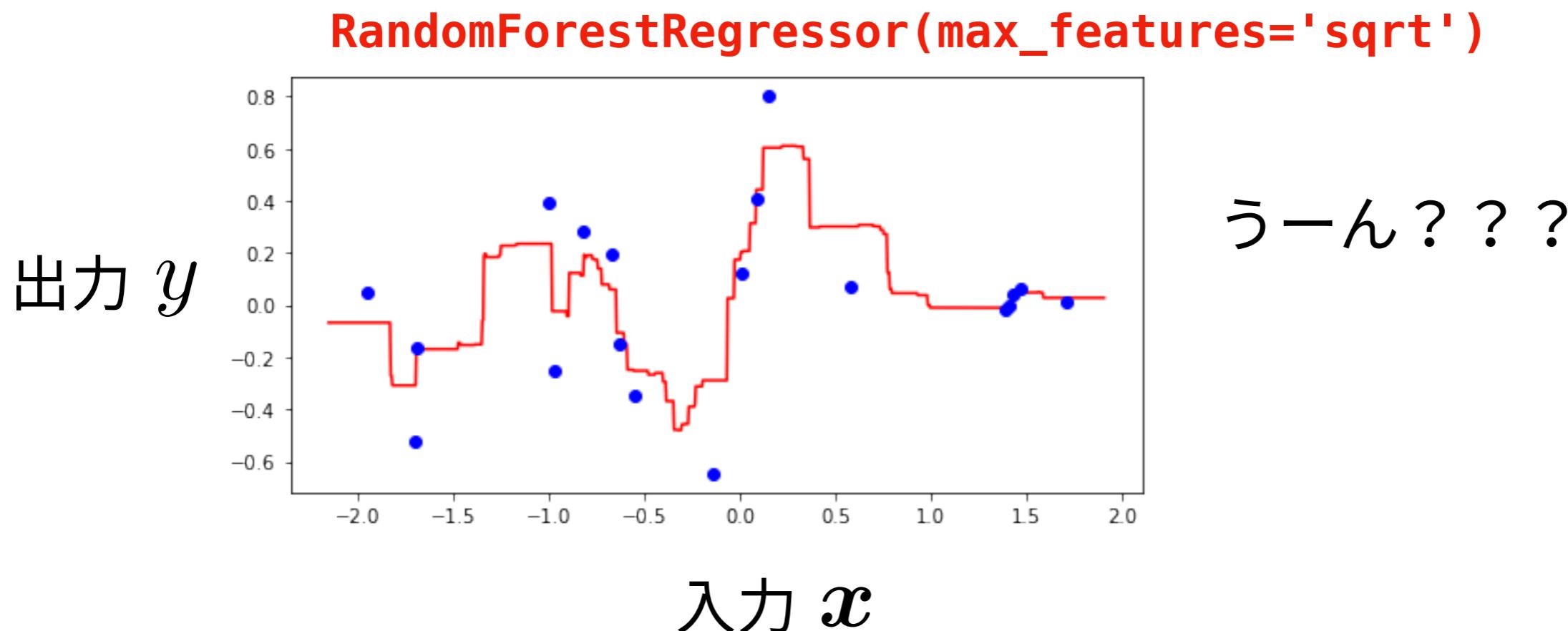
GaussianProcessRegressor(kernel=Matern(length_scale=100.0, nu=0.2))



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

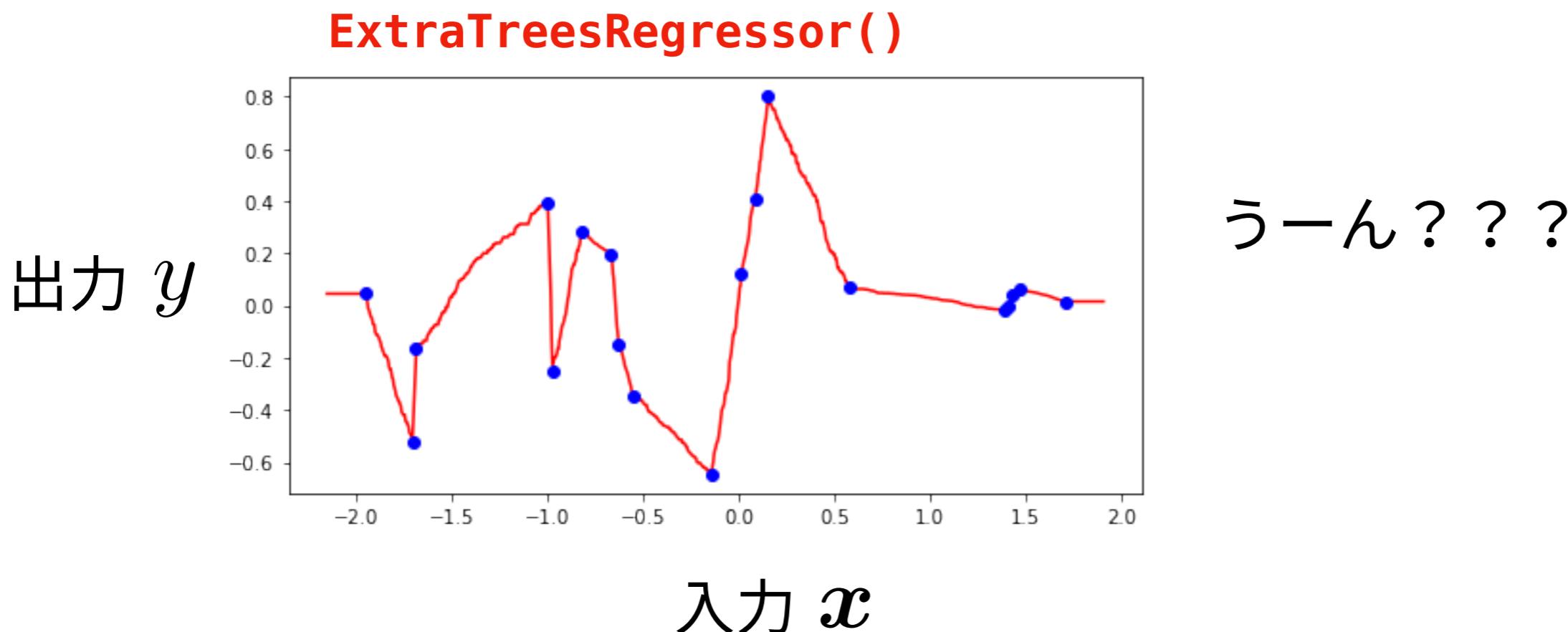
このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？

入力 x

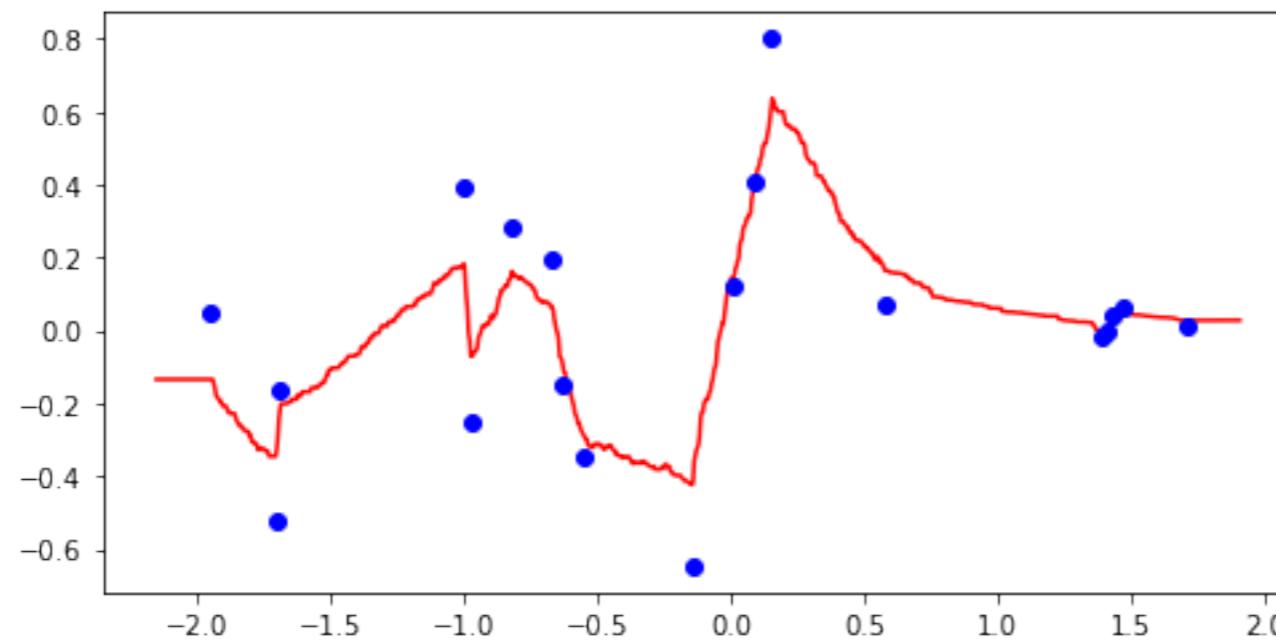
- 元素組成比
- 実験条件

機械学習モデル

出力 y

- 収率

出力 y



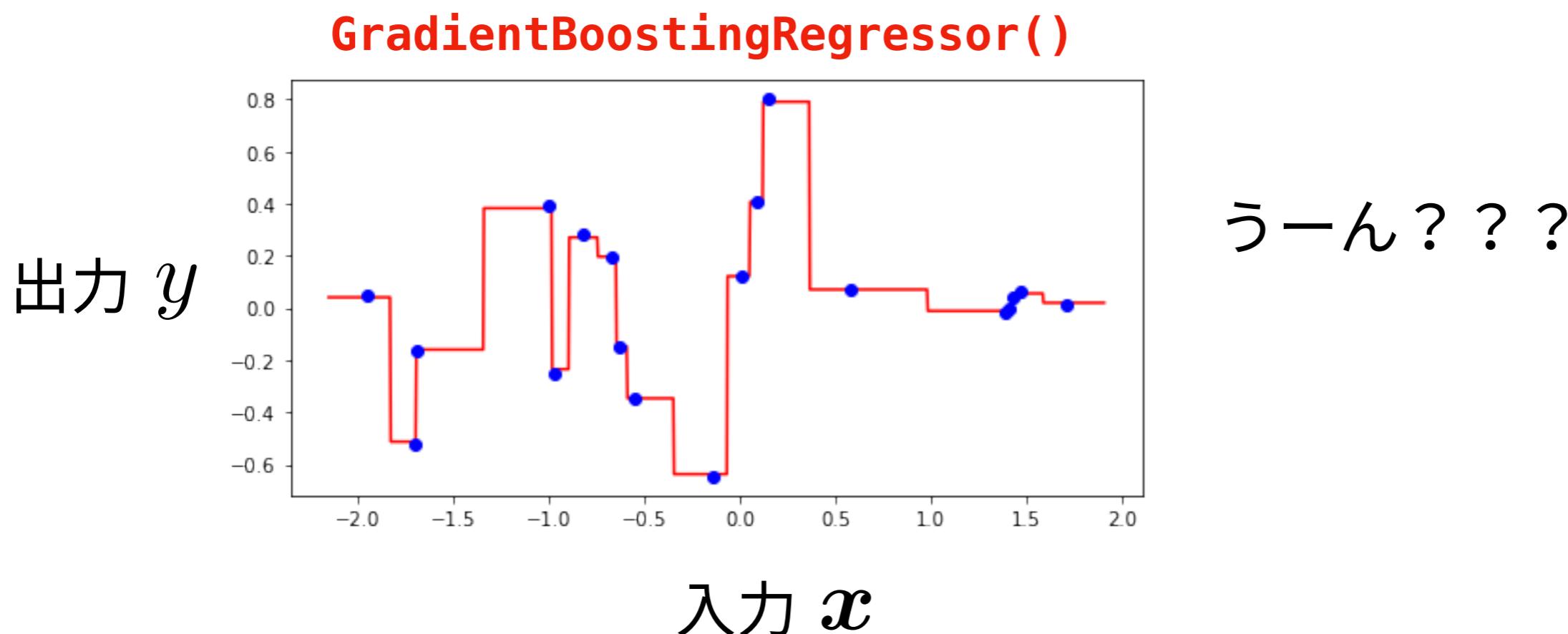
入力 x

sklearnのデフォルト
はoffなので注意

うーん？？？

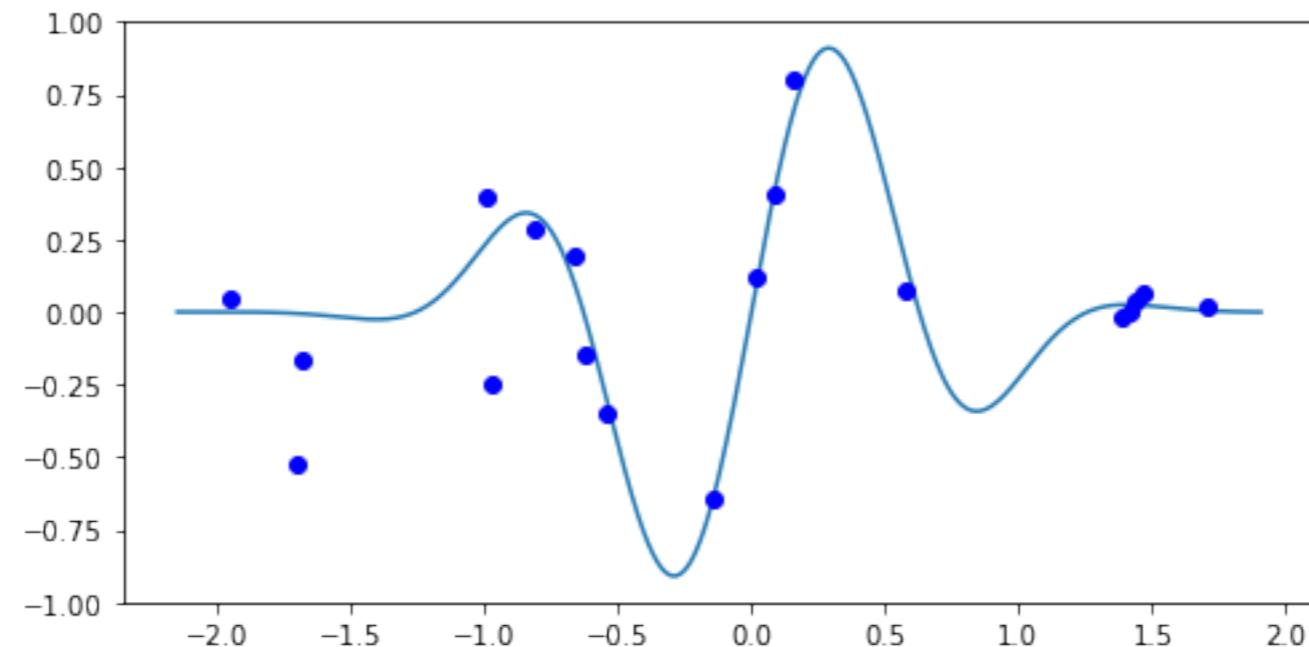
このとき機械学習に何が必要か

この表データで機械学習モデルを訓練し予測すれば良い？



この例は実は「真のモデル+ノイズ」の人工データ

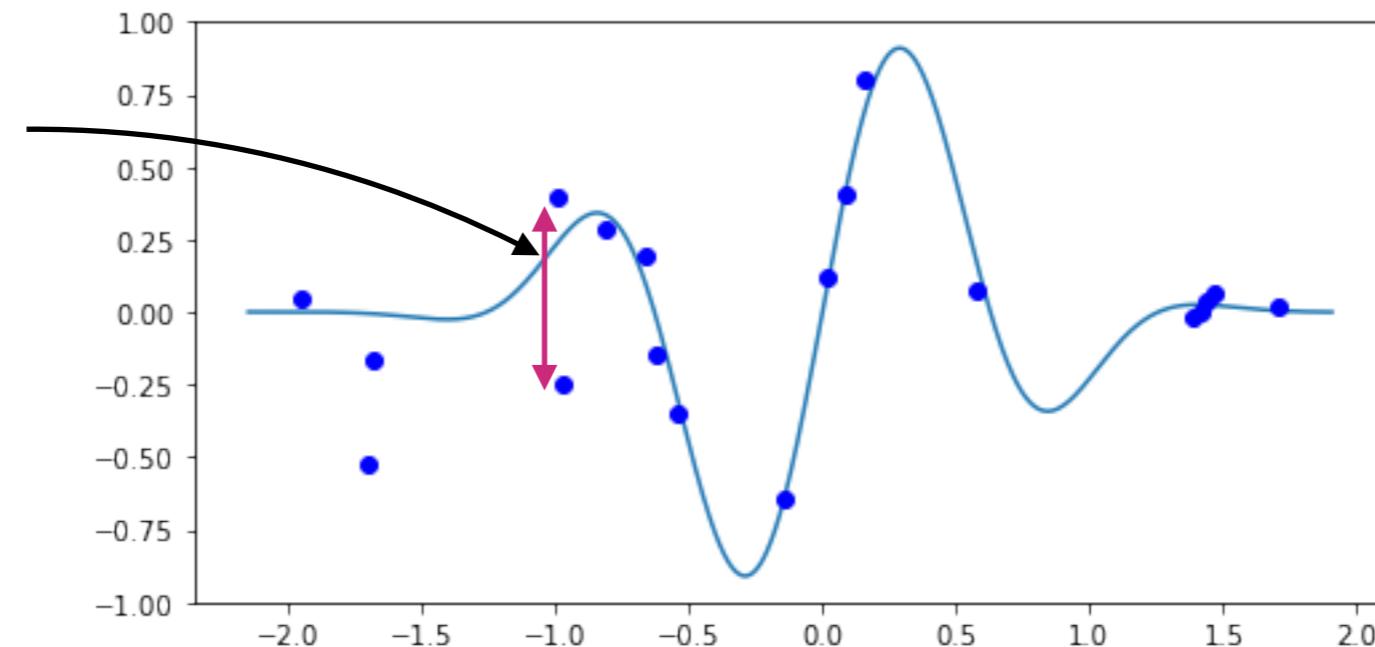
現実の実測データ(ましてや報告データ)には色々な落とし穴が！



この例は実は「真のモデル+ノイズ」の人工データ

現実の実測データ(ましてや報告データ)には色々な落とし穴が！

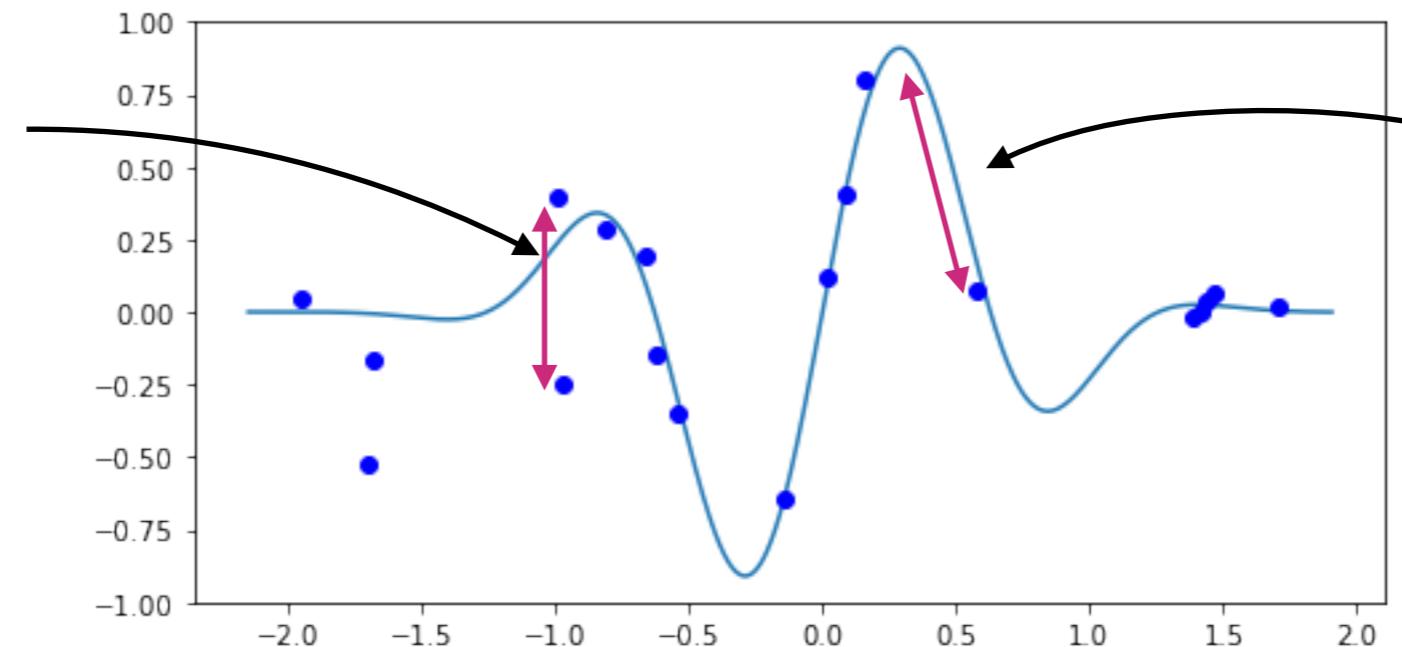
近い正解例が
inconsistent
(教師ノイズ)



この例は実は「真のモデル+ノイズ」の人工データ

現実の実測データ(ましてや報告データ)には色々な落とし穴が！

近い正解例が
inconsistent
(教師ノイズ)

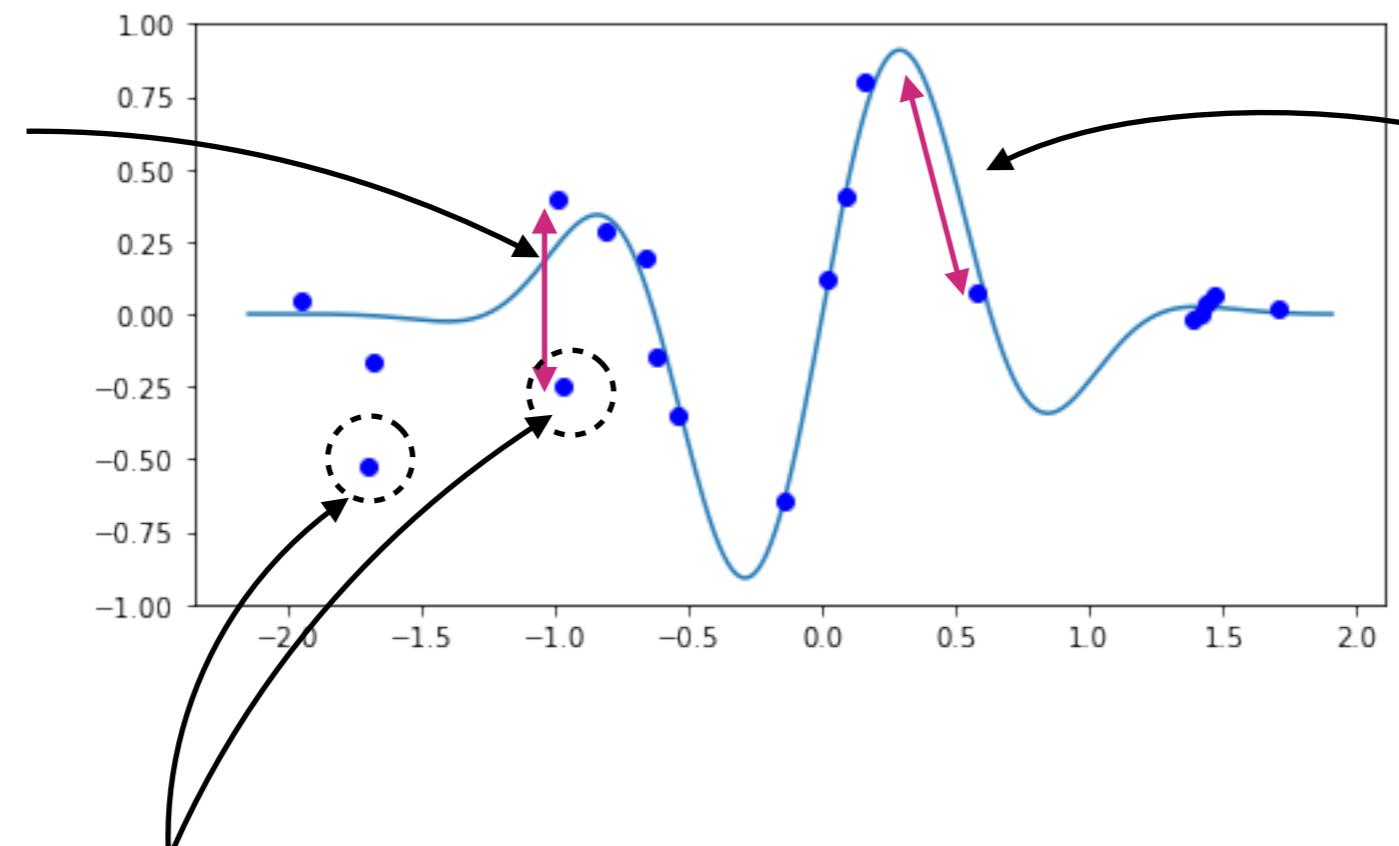


xの少しの違いで
yに急峻な変化
(Cliffs)

この例は実は「真のモデル+ノイズ」の人工データ

現実の実測データ(ましてや報告データ)には色々な落とし穴が！

近い正解例が
inconsistent
(教師ノイズ)

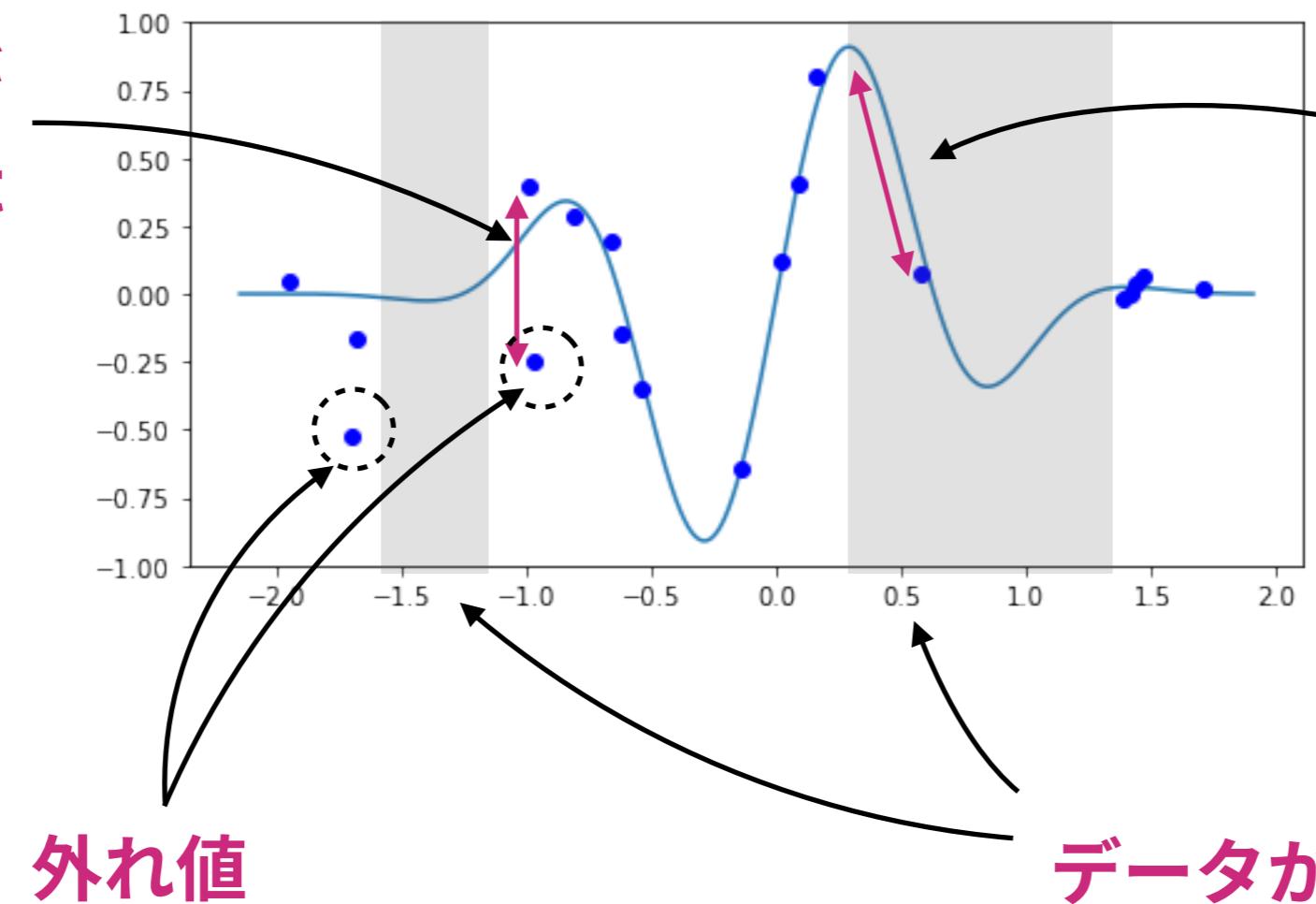


xの少しの違いで
yに急峻な変化
(Cliffs)

この例は実は「真のモデル+ノイズ」の人工データ

現実の実測データ(ましてや報告データ)には色々な落とし穴が！

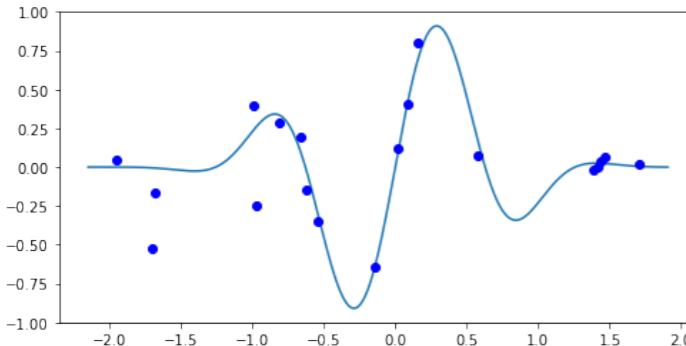
近い正解例が
inconsistent
(教師ノイズ)



xの少しの違いで
yに急峻な変化
(Cliffs)

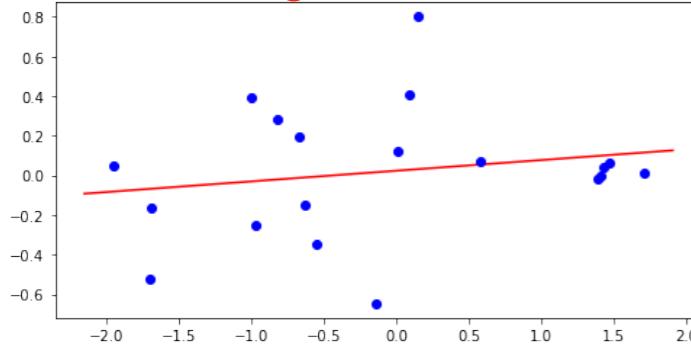
データがない/少ない領域

この例は実は「真のモデル+ノイズ」の人工データ

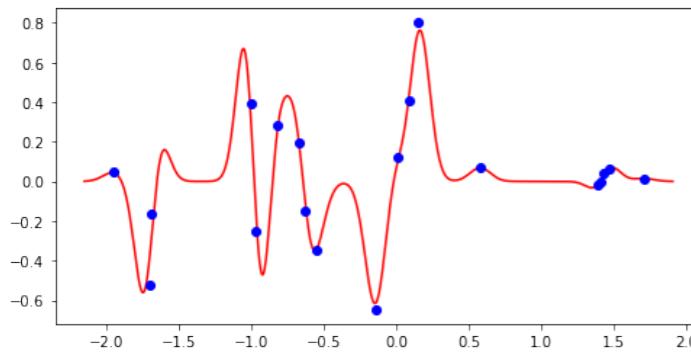


← “真の”モデル

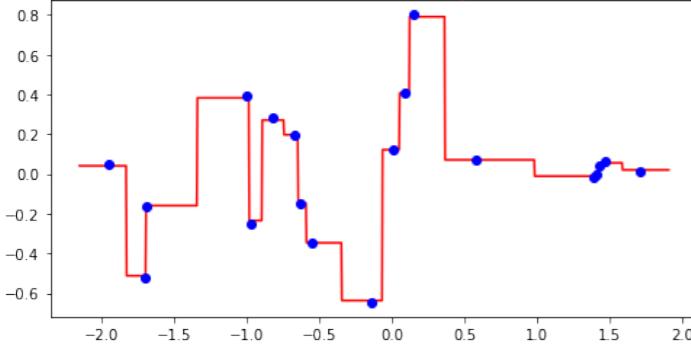
Linear Regression



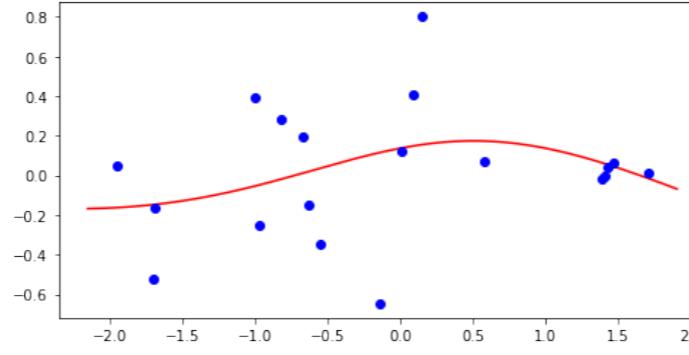
Kernel Ridge (RBF)



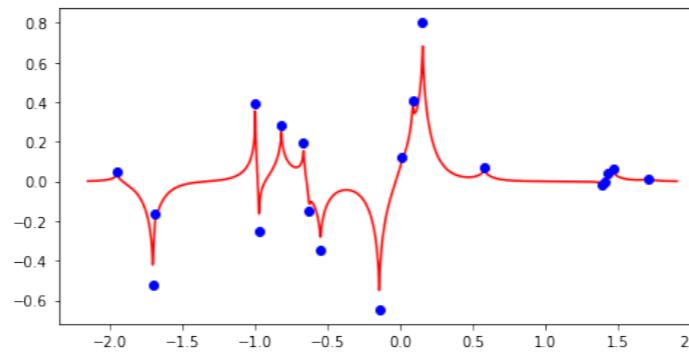
Gradient Boosting



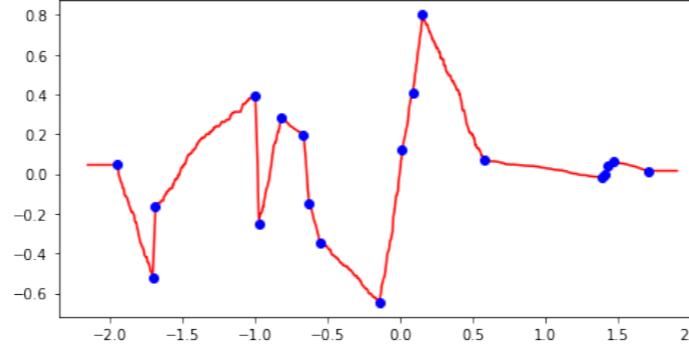
Neural Networks (Tanh)



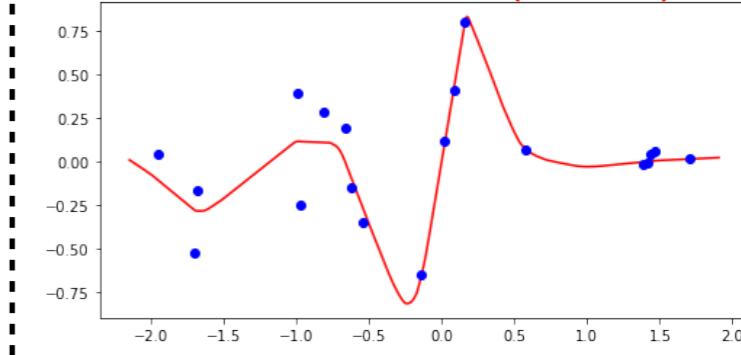
Kernel Ridge (Laplacian)



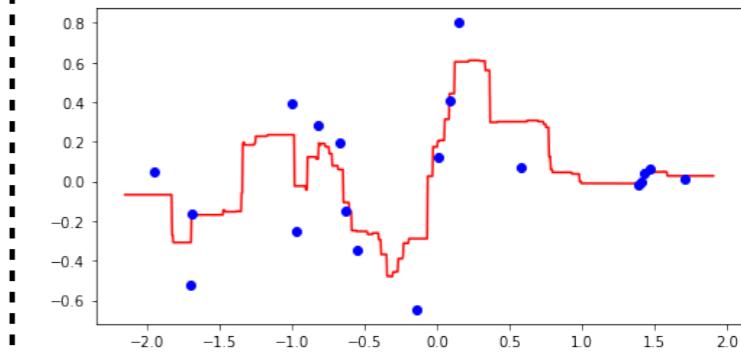
Extra Trees (no bootstrap)



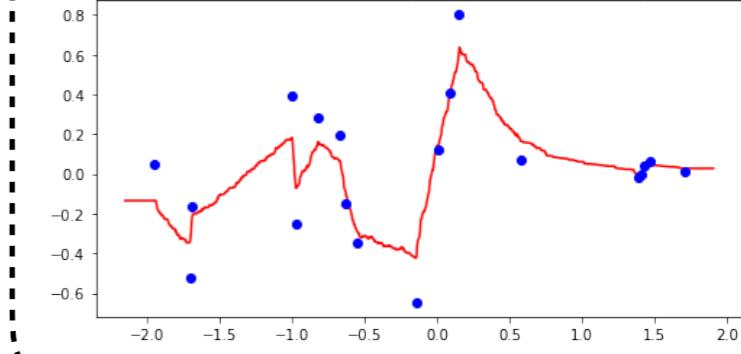
Neural Networks (ReLU)



Random Forest



Extra Trees (bootstrap)



Rashomon効果：一体どれを信じればいいんじやい！

🤔 機械学習モデルや訓練データが変われば予測は変わる

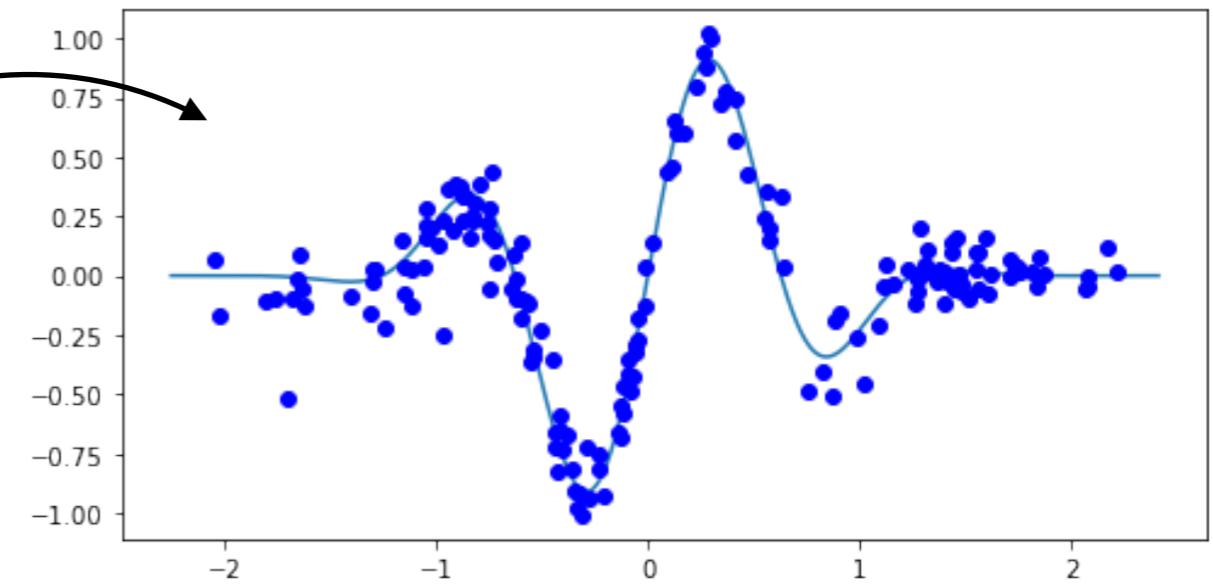
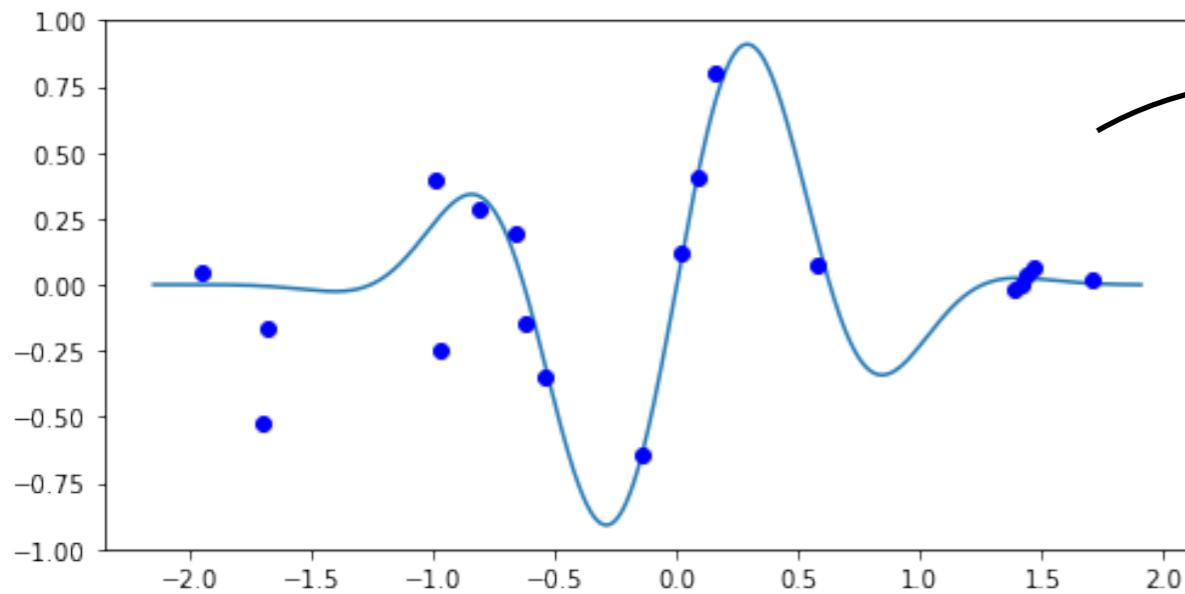
(モデルとデータの数だけ予測もある)

- Cross-validation精度はほぼ同等のモデルが無数にあり得る
- 同じモデルでもHyperparameterが違えばかなり違い得る
- 現実では真のモデルは分からぬため良し悪しの判断は困難
- 実際には「Fittingを非常に高次元でやる」ので非直感的なことが色々起こる(私の1次元の例の作為性にご注意を!)

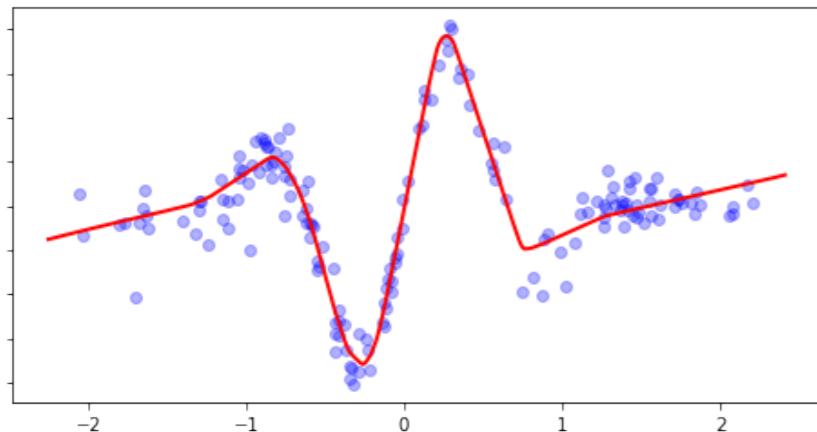
MLがどういう技術なのか**MLの特性と限界**を正しく把握する
予測結果を当該分野の知識に照らして注意深く検証・解釈する

ちなみにサンプル数が十分大きければわりとどれでもOK

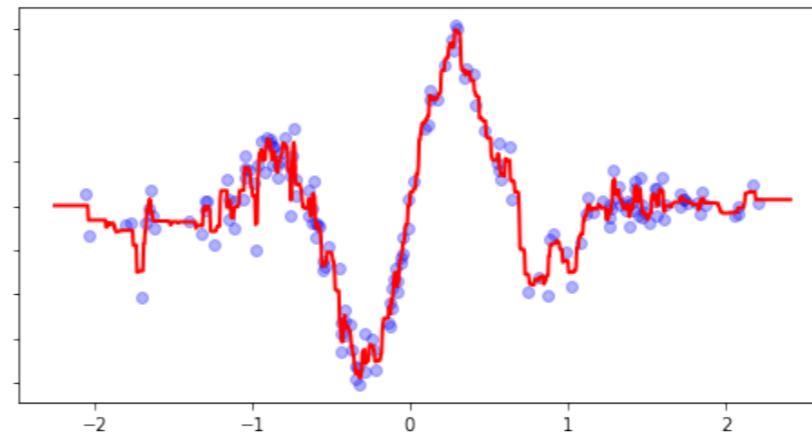
サンプル数が十分多ければ例外的データ点は統計的に相殺される
→ ただし高次元であるほど指数的な数が必要で非現実的な期待



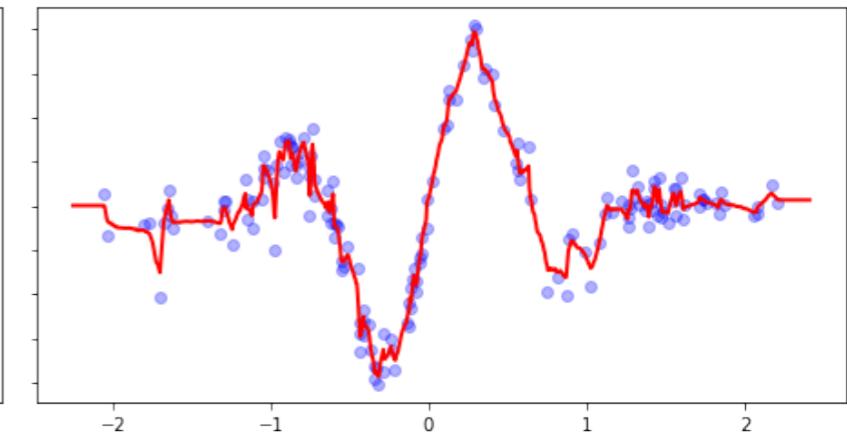
Neural Networks (ReLU)



Random Forest

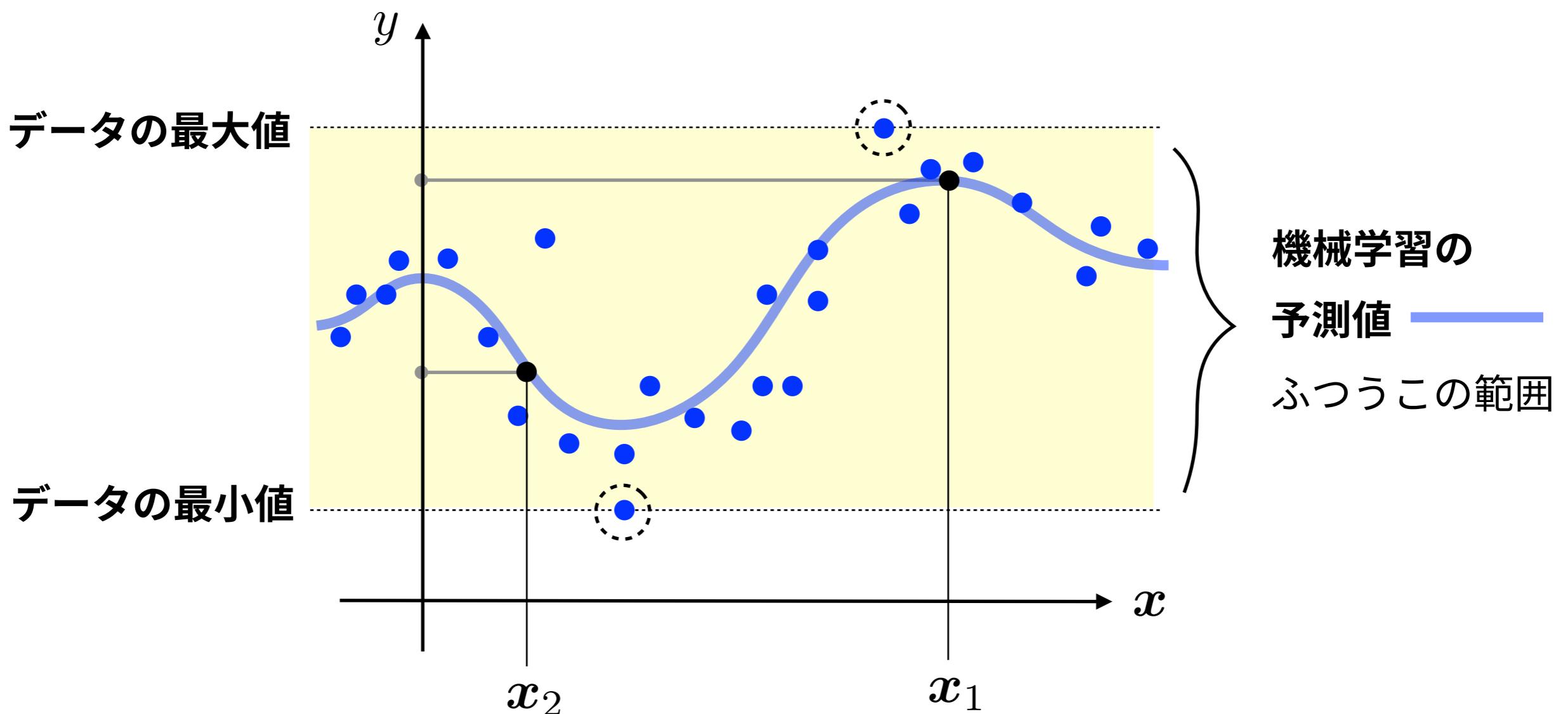


Extra Trees (bootstrap)



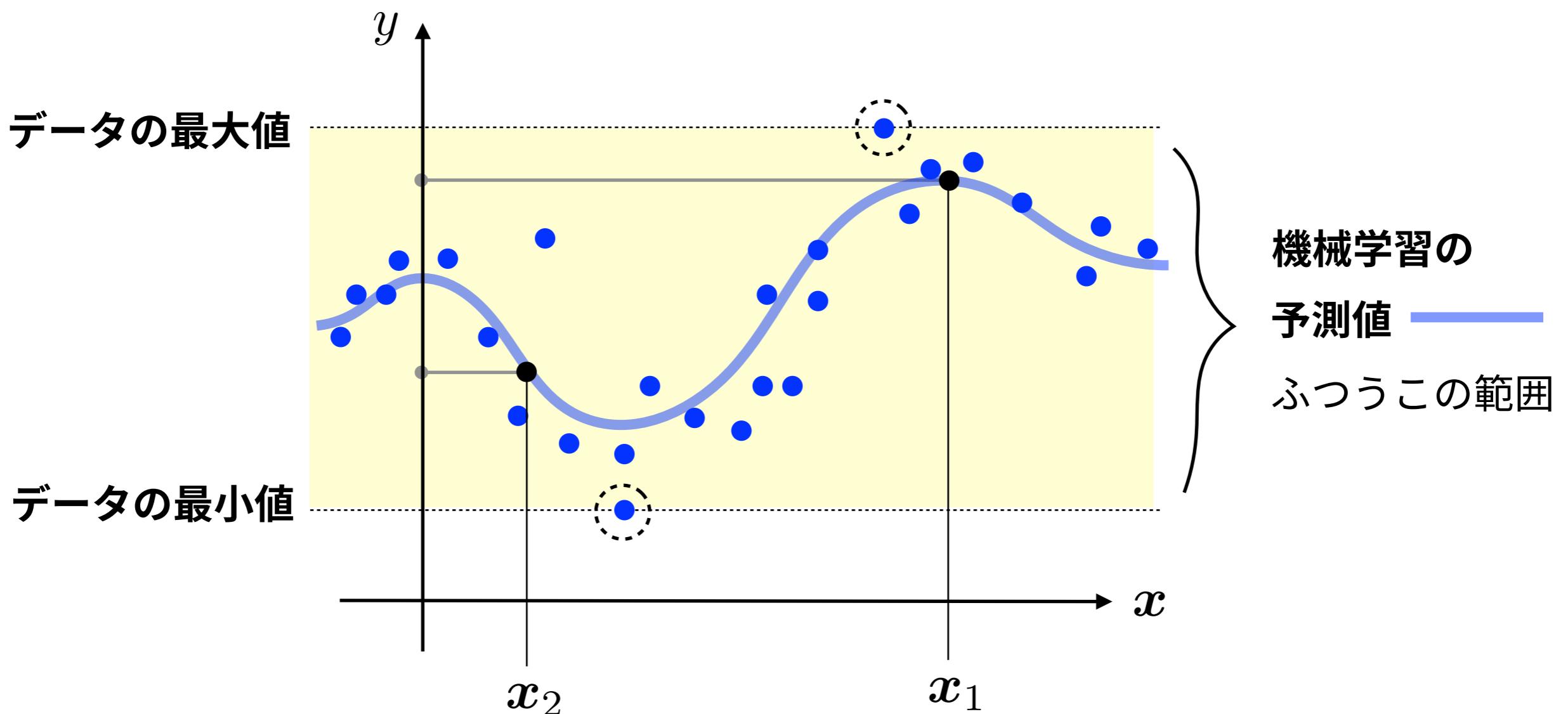
機械学習の予測値は訓練データの最良値をふつう超えない

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



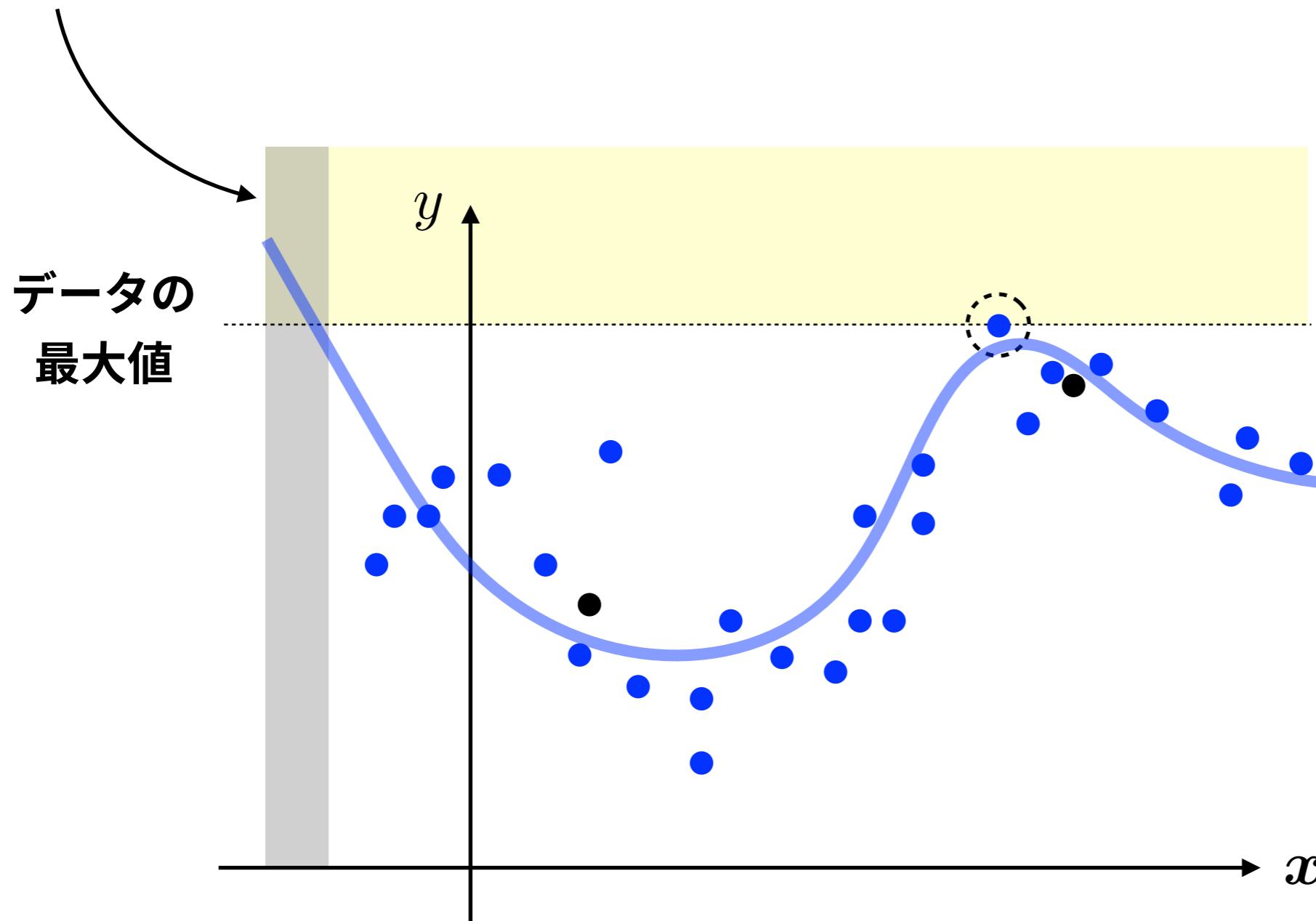
機械学習の予測値は訓練データの最良値をふつう超えない

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



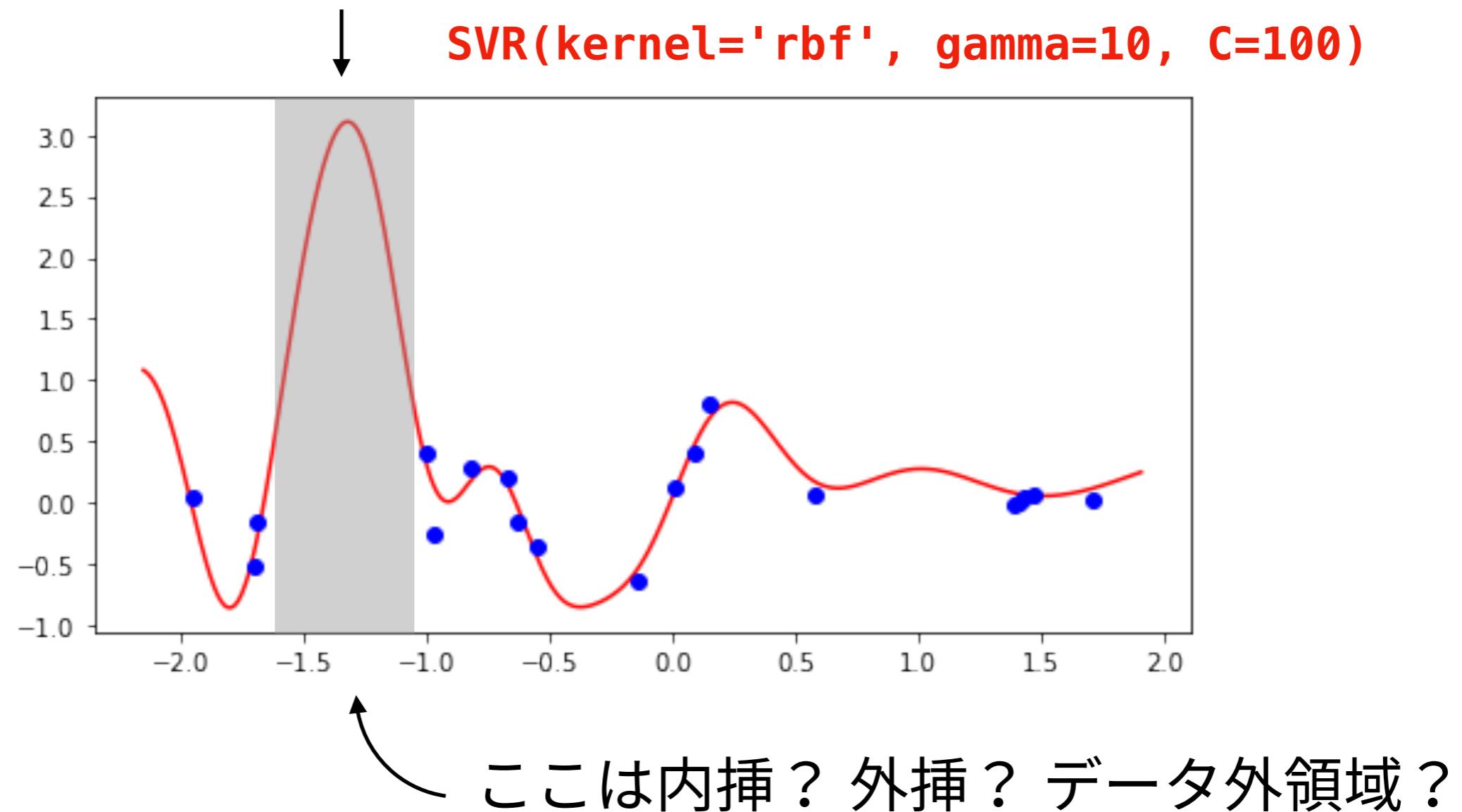
機械学習の予測値は訓練データの最良値をふつう超えない

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



「データ外領域での意図しない外挿」リスク

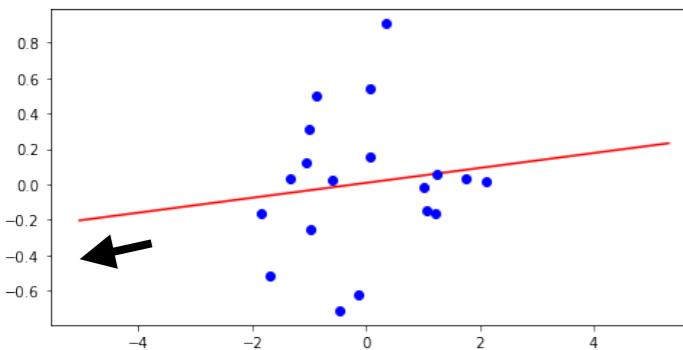
1次元の例では「内挿」か「外挿」か分かる感じがしてしまうが
一般的の高次元ではこの判別すら直感的ではないことに注意



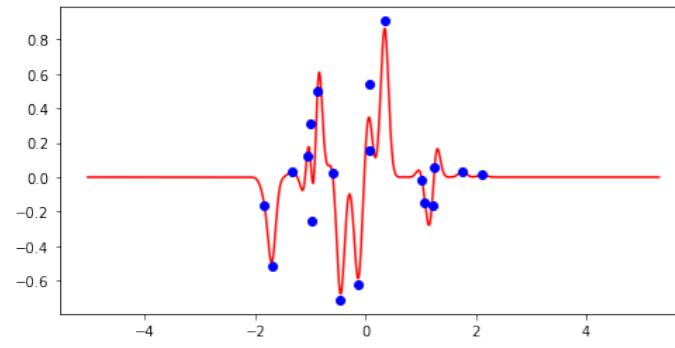
「データ外領域での意図しない外挿」 リスクは手法に依存

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

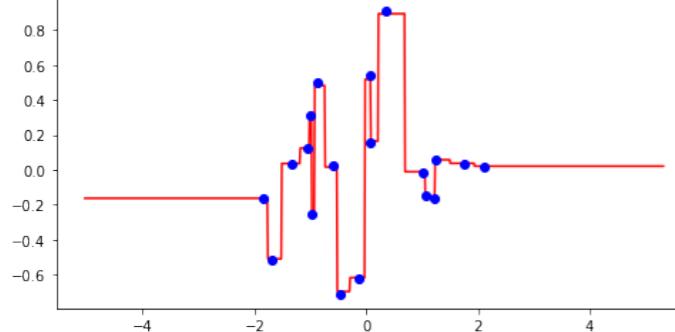
Linear Regression



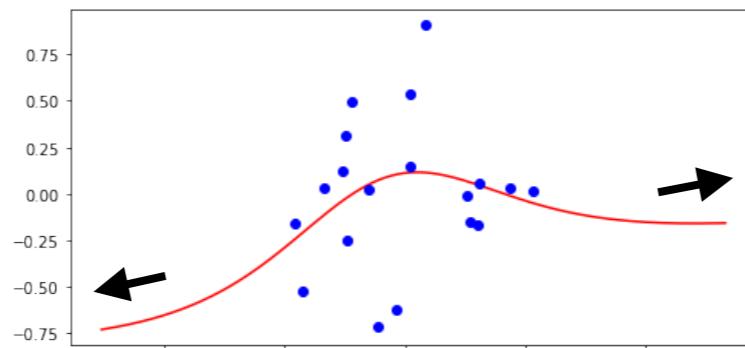
Kernel Ridge (RBF)



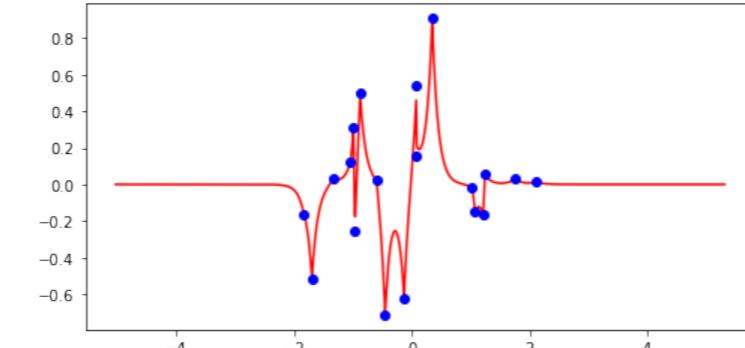
Gradient Boosting



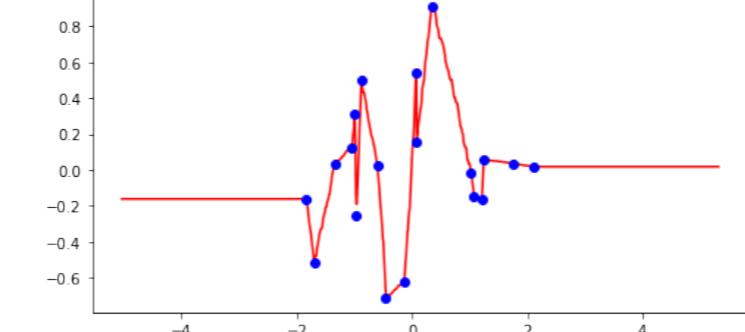
Neural Networks (Tanh)



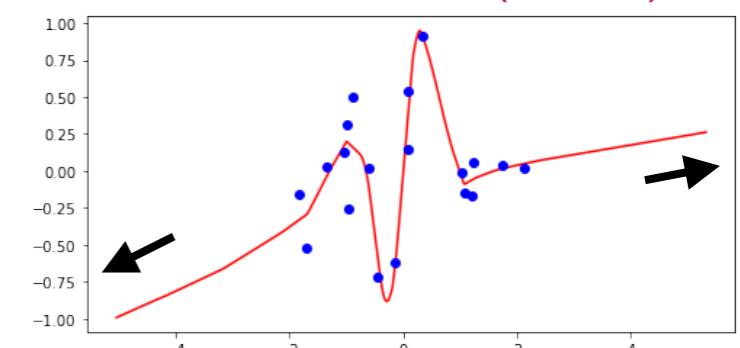
Kernel Ridge (Laplacian)



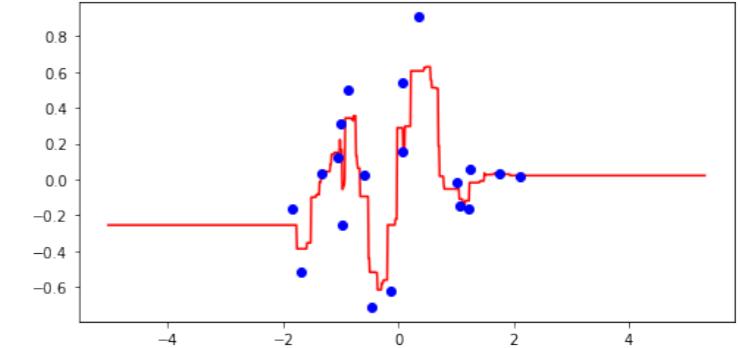
Extra Trees (no bootstrap)



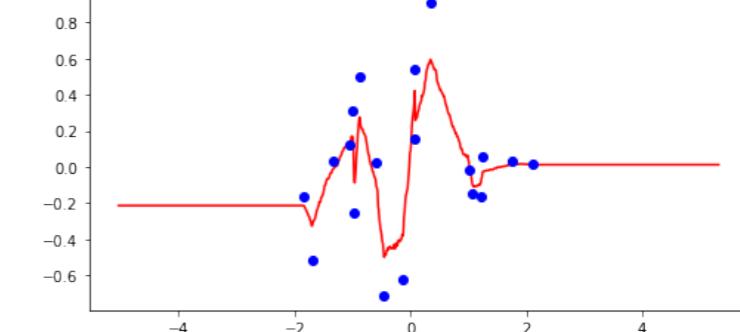
Neural Networks (ReLU)



Random Forest



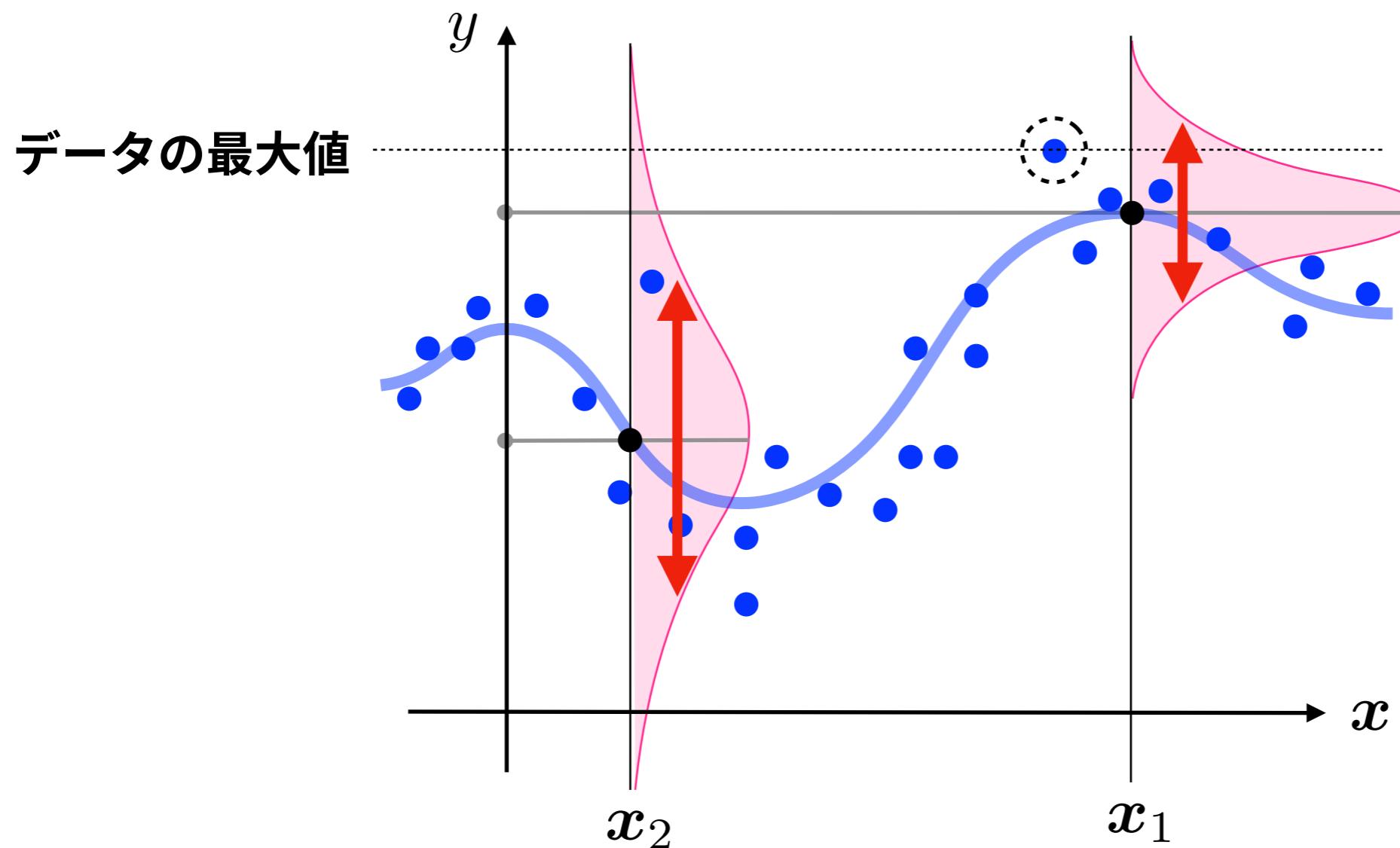
Extra Trees (bootstrap)



予測値だけではなくその予測分散(確度)も考えるのが重要

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

e.g. 「収率予測値は 20.2 ± 15.1 」 vs 「収率予測値は 20.2 ± 2.5 」

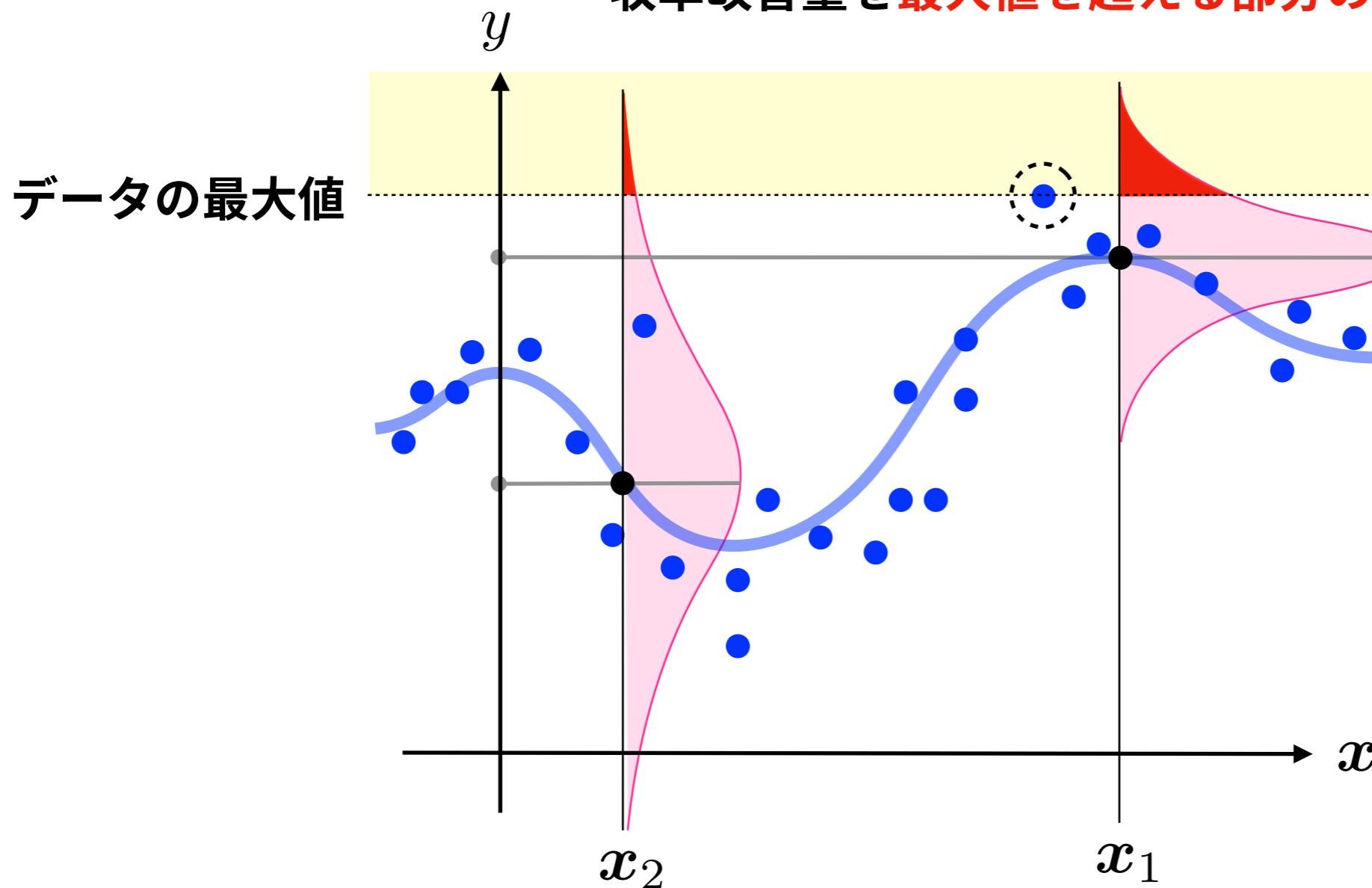


探索する際は予測値 자체を指標としない

探索の目的では期待改善(EI)や信頼区間の上限などを指標に

期待改善(EI) = 収率改善量の期待値

収率改善量を最大値を超える部分の確率の重みで積分



文献から集めた実際の実験データ報告を使う

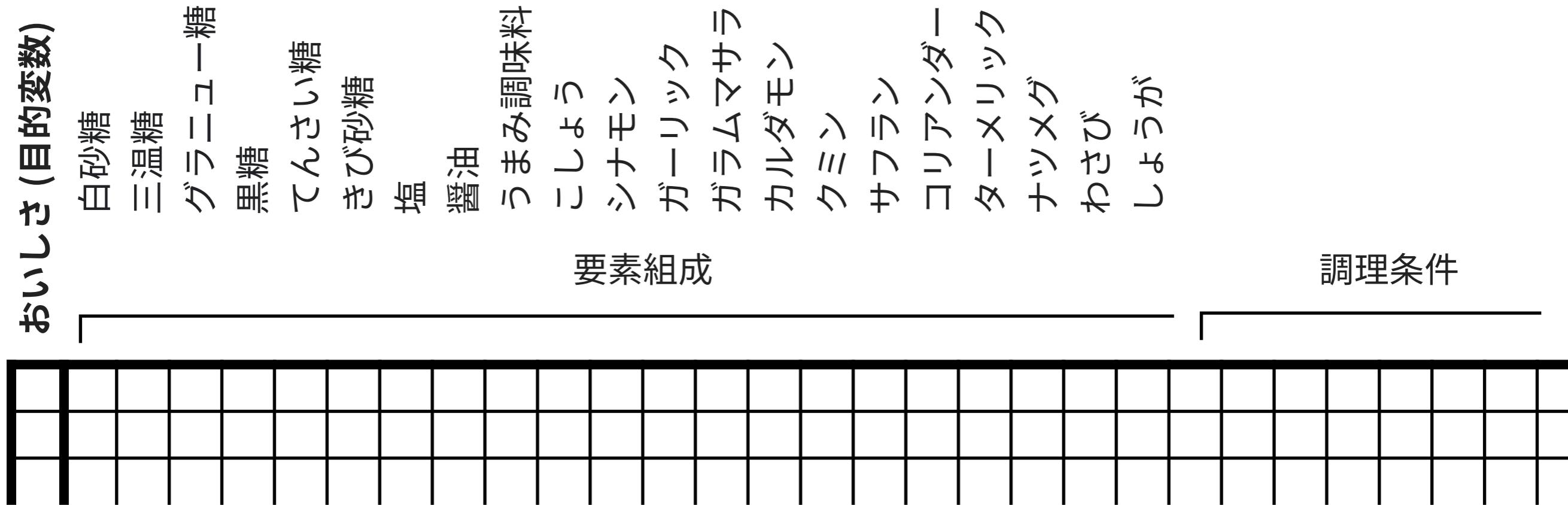
対象はメタンの酸化カップリング反応、目的変数はC₂収率

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



Nr of public	A	B	C	N	O	R	S	T	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM
	Cation 1	Cation 1	Anion 1	Anion 1	Promotor	Support	Support	Preparati	Temperat	p(CH ₄), bar	p(O ₂), bar	p(CH ₄)/p(O ₂)	p total,	Contact time, s	X(O ₂), %	X(CH ₄), %	S(CO _x), %	S(C ₂ ⁻), %	S(C ₂ ⁻), %	S(C ₂), %	Y(C ₂), %		
1	Mn	9.2				AI	90.8	Impregnat	1073	0.40	0.08	4.8	1.0		0.04		11.0				45.5	5.0	
20	Li	30.3				n.a.		993	0.08	0.04	2.0	1.0		5.30	85.0	38.0					50.0	19.0	
21	Mg	66.7	S	33.3				Impregnat	1019	0.65	0.08	8.1	1.0		1.40	39.0	4.0		23.0	41.0	64.0	2.6	
22	Mg	55.0	S	45.0				Impregnat	1017	0.66	0.08	8.3	1.0		3.00	65.0	10.0		27.0	40.0	67.0	6.7	
23	Na	7.0	S	60.0				Impregnat	1017	0.64	0.08	8.0	1.0		0.19	39.0	3.0		23.0	19.0	42.0	1.3	
75	Pb	20.0				AI	80.0	n.a.	1030	0.96	0.05	19.2	1.0		0.40	100.0	6.8		17.6	32.8	50.4	3.4	
76	Pb	20.0				Si	80.0	n.a.	1103	0.96	0.05	19.2	1.0		0.55	44.1	18.7		18.7	20.5	39.2	7.3	
486	K	3.0	Cl	3.0	Cl	AI	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		30.8				33.6	10.3	
487	Li	3.0	Cl	3.0	Cl	AI	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		2.2				76.9	1.7	
488	Ba	3.0	Cl	6.0	Cl	AI	82.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		32.1				32.1	10.3	
489	Na	3.0	Cl	3.0	Cl	AI	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		35.0				33.5	11.7	
490	Cs	3.0	Cl	3.0	Cl	AI	85.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		30.2				24.3	7.3	
491	Ag	18.0			Cl	AI	82.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		20.4				0.0	0.0	
492	Ag	18.0	C	41.0	Cl			Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.8				0.3	0.1	
493	Pr	5.0			Cl	AI	86.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.6				0.2	0.1	
494	Pr	1.0			Cl	AI	90.5	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.1				0.0	0.0	
495	Bi	1.0			Cl	AI	81.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		23.4				1.3	0.3	
496	Ba	1.0			Cl	AI	81.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		27.8				0.7	0.2	
497	Ba	5.0			Cl	AI	77.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		26.0				1.7	0.4	
498	K	3.0			Cl	AI	79.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		17.2				23.3	4.0	
499	Ba	3.0	Cl	6.0	Cl	AI	82.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		15.8				28.0	4.4	
500	Ba	3.0	Cl	6.0	Cl	AI	73.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		27.1				30.4	8.2	
501	Ca	1.0	Cl	2.0	Cl	AI	79.0	Impregnat	973	0.10	0.05	2.0	1.0		1.50		15.8				25.4	4.0	
502	Ag	18.0			Cl	AI	82.0	Therm.dec	973	0.10	0.05	2.0	1.0		1.50		5.0				0.0	0.0	
503	Ba	3.0	Cl	6.0	Cl	AI	73.0	Therm.dec	973	0.10	0.05	2.0	1.0		1.50		17.2				25.4	4.4	
504	Ba	3.0	Cl	6.0	Cl			Therm.dec	973	0.10	0.05	2.0	1.0		1.50		26.7				15.3	4.1	
505	Ba	3.0	Cl	6.0	Cl	AI	73.0	Therm.dec	973	0.10	0.05	2.0	1.0		1.50		21.3				30.4	6.5	
506	Sr	3.0	Cl	6.0	Cl	AI	91.0	Impregnat	1023	0.10	0.05	2.0	1.0		1.50		30.3				56.0	17.0	
507	Ba	28.0	C	28.0	Cl	AI	44.0	Impregnat	1023	0.10	0.05	2.0	1.0		1.50		43.2				41.8	18.1	
508	Ba	2.0	Cl	2.0	Cl	AI	21.0	Impregnat	1010	0.00	0.05	1.0	1.0		0.10		17.0				11.0	10.0	

特徴量の設計：触媒の効果的な特徴表現を考える



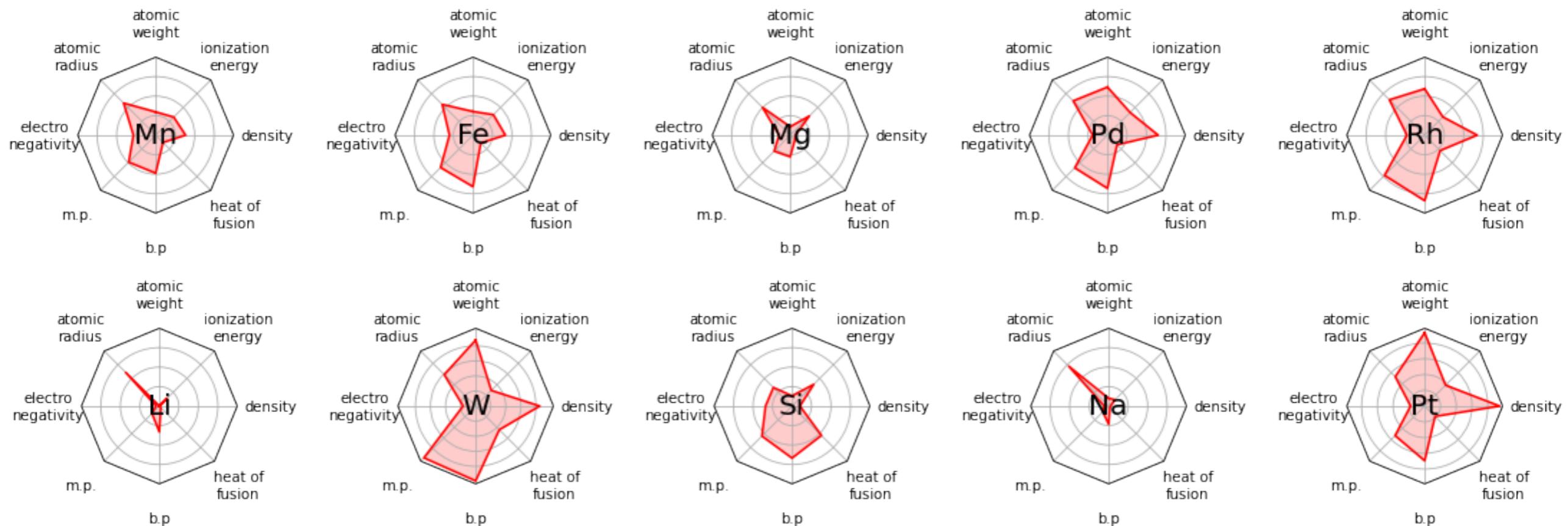
- 白砂糖～きび砂糖まではショ糖で「甘味」、醤油には「塩分」が含まれる、など、それぞれの要素の「個性」は全く考慮されない。
- 訓練データに含まれない要素が入ると予測に反映できない。
- 要素ごとの頻度がベキ乗則的であり非常に大きな偏りがある。
- 要素数が多く報告例に要素のオーバラップが少ない。

元素の個性を元素記述子ベクトルで表す入力表現

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

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

Element	Descriptors			
	1	2	...	p
A	A ₁	A ₂	...	A _p
B	B ₁	B ₂	...	B _p
C	C ₁	C ₂	...	C _p
D	D ₁	D ₂	...	D _p
E	E ₁	E ₂	...	E _p



Sorted Weighted Elemental Descriptors (SWED)

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

Catalyst	1 st feature				2 nd feature				...
	1	2	...	p	1	2	...	p	
Cat-ABC1	90 × A ₁	90 × A ₂	...	90 × A _p	6 × C ₁	6 × C ₂	...	6 × C _p	...
Cat-BDE1	80 × D ₁	80 × D ₂	...	80 × D _p	11 × B ₁	11 × B ₂	...	11 × B _p	...
Cat-AE1	75 × A ₁	75 × A ₂	...	75 × A _p	25 × E ₁	25 × E ₂	...	25 × E _p	...
Cat-AE2	80 × A ₁	80 × A ₂	...	80 × A _p	20 × E ₁	20 × E ₂	...	20 × E _p	...
Cat-ABCDE1	90 × E ₁	90 × E ₂	...	90 × E _p	15 × C ₁	15 × C ₂	...	15 × C _p	...

Element	Descriptors			
	1	2	...	p
A	A ₁	A ₂	...	A _p
B	B ₁	B ₂	...	B _p
C	C ₁	C ₂	...	C _p
D	D ₁	D ₂	...	D _p
E	E ₁	E ₂	...	E _p

- アグレッシブな探索向けにSWEDからの組成復元法も開発
- 組成比をかけないとシンボルとして扱うのと等価になる
- 組成比をかける操作は交互作用やゲーティングとみなせる
- 色々な行列分解、アイチソン幾何の考慮、…も試したが×
- 組合せ集合やグラフの機械学習の技術で解く方法を開発中

元素記述子を反映した入力表現

1. **Conventional:** 組成+実験条件
2. **Proposed(Exploitative):** 組成+SWED(8記述子)+実験条件
3. **Proposed(Explorative):** SWED(3 or 8記述子)+実験条件

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

^[a] The electronegativity, density, and ΔH_{fus} were used as descriptors.

学習には「知識の利用」と「探索」のトレードオフが伴う

新しいことを「学ぶ」際の最も基本的なトレードオフ

1. 今まで学んだことの「利用」

今までのデータの機械学習に基づく予測の活用

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

2. 今までに学んでないことの「探索」

= 新しい経験、新しい知識の吸収、知識の拡充

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

文献データには様々な問題があり「探索」がより重要

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

LETTER

12 SEPTEMBER 2019 | VOL 573 | NATURE | 251

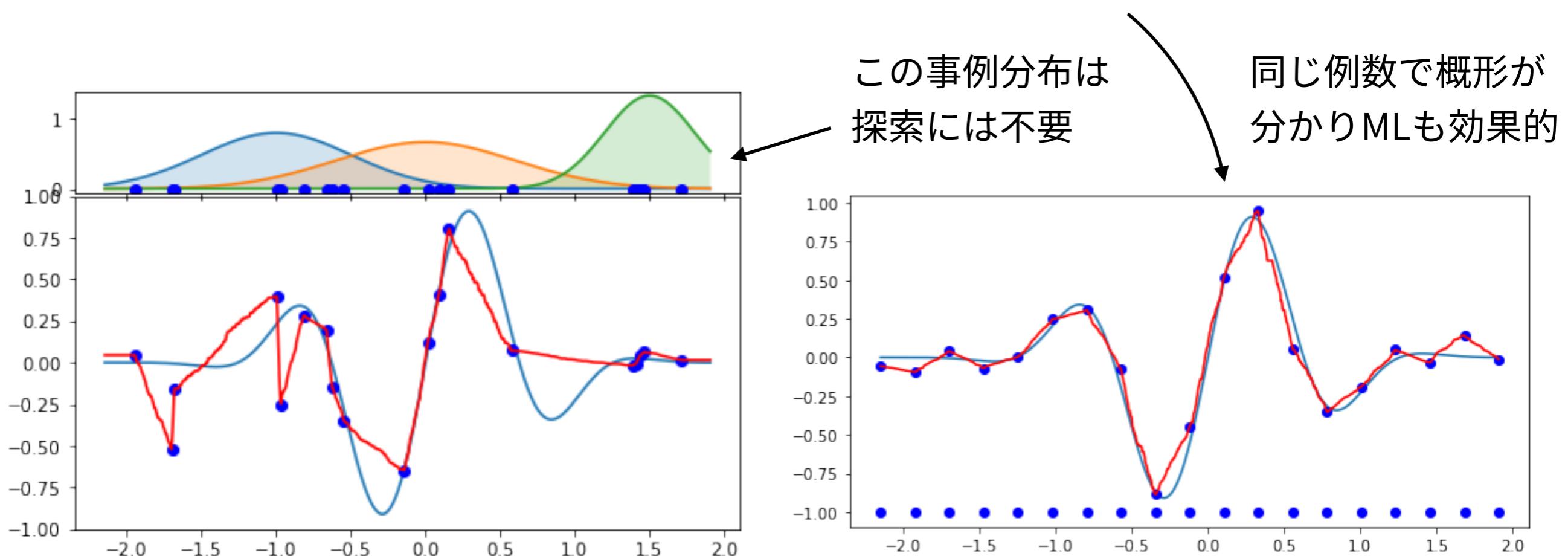
<https://doi.org/10.1038/s41586-019-1540-5>

Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis

Xiwen Jia¹, Allyson Lynch¹, Yuheng Huang¹, Matthew Danielson¹, Immaculate Lang'at¹, Alexander Milder¹, Aaron E. Ruby¹, Hao Wang¹, Sorelle A. Friedler^{2*}, Alexander J. Norquist^{1*} & Joshua Schrier^{1,3*}

実験計画と機械学習

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



実験計画におけるフィッシャーの三原則

実験計画におけるフィッシャーの三原則

1. 反復 (replication)

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

実験計画におけるフィッシャーの三原則

1. 反復 (replication)

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

2. 無作為化 (randomization)

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

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

実験計画におけるフィッシャーの三原則

1. 反復 (replication)

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

2. 無作為化 (randomization)

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

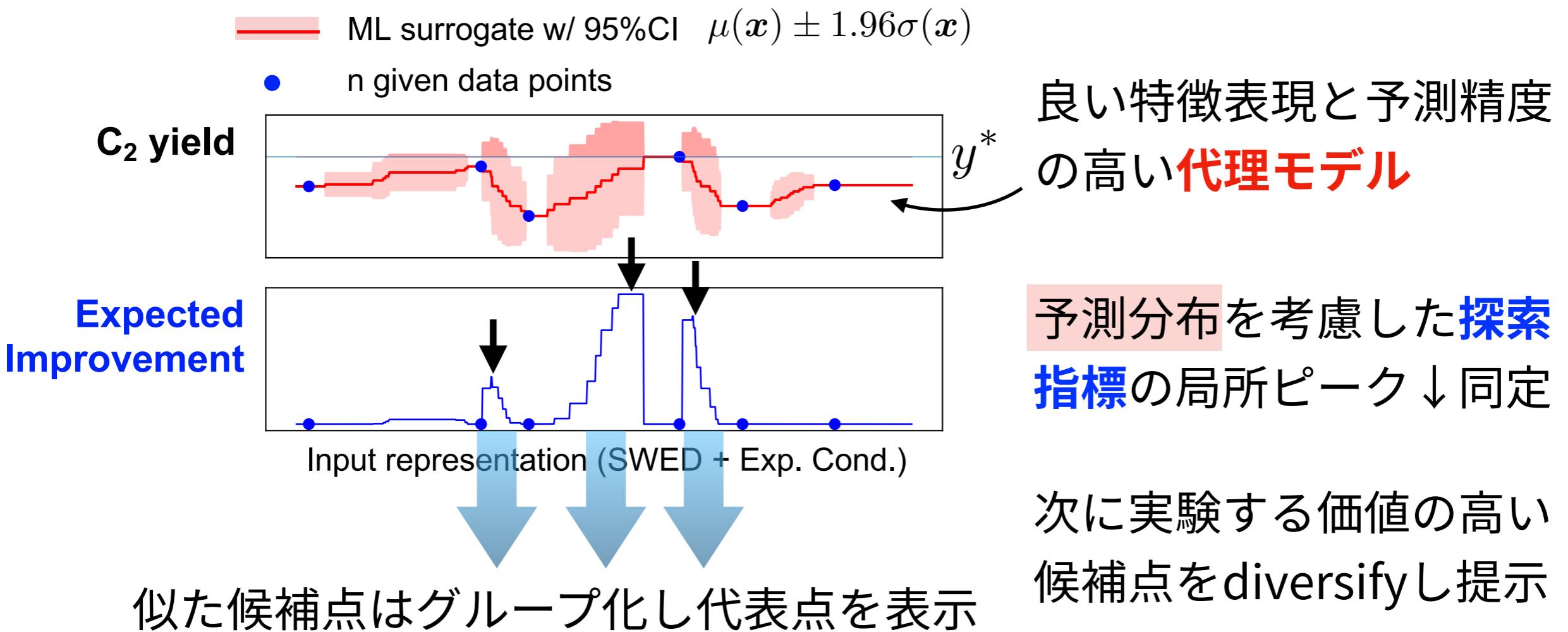
(c.f. 結果がよかった条件まわりで多めに試したいのは理解できるが探索が目的ならランダム実験条件のほうが良い)

3. 局所管理 (local control)

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

(c.f. 実験条件の最適化は諦めて固定し組成だけふる実験)

実験計画と機械学習

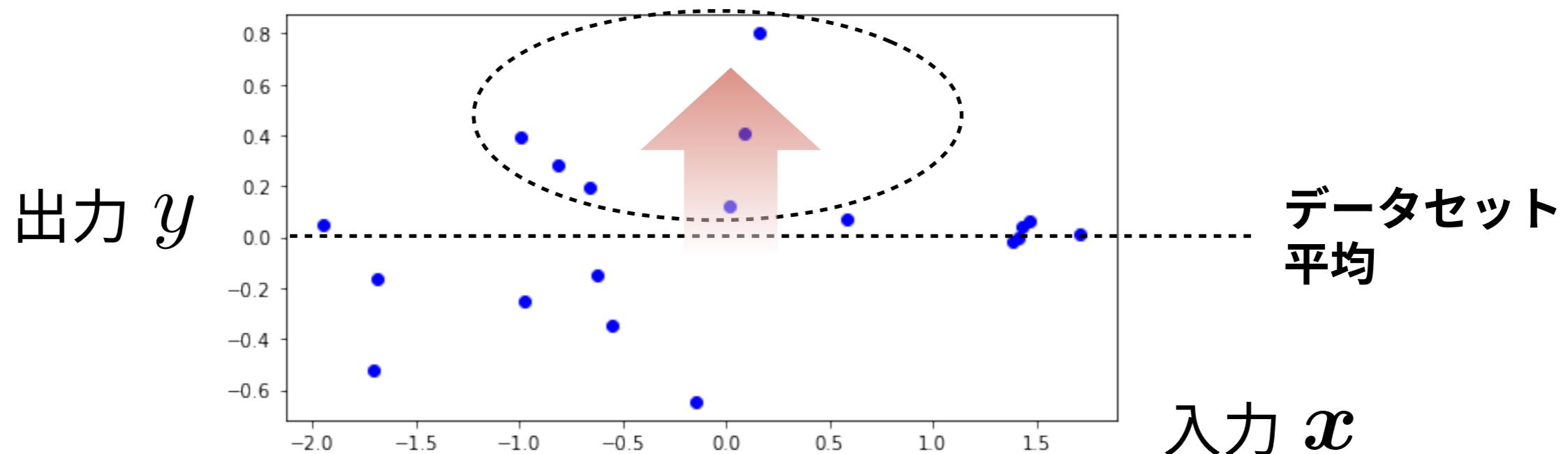


MLに入力するデータにない傾向は原理上予測できないため
アルゴリズムの詳細よりも **データの収集計画 (実験計画)、適用範囲の理解、品質保証** が成功の鍵であることをいつも心に

機械学習モデルがなぜその予測をしたかの要因分析



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

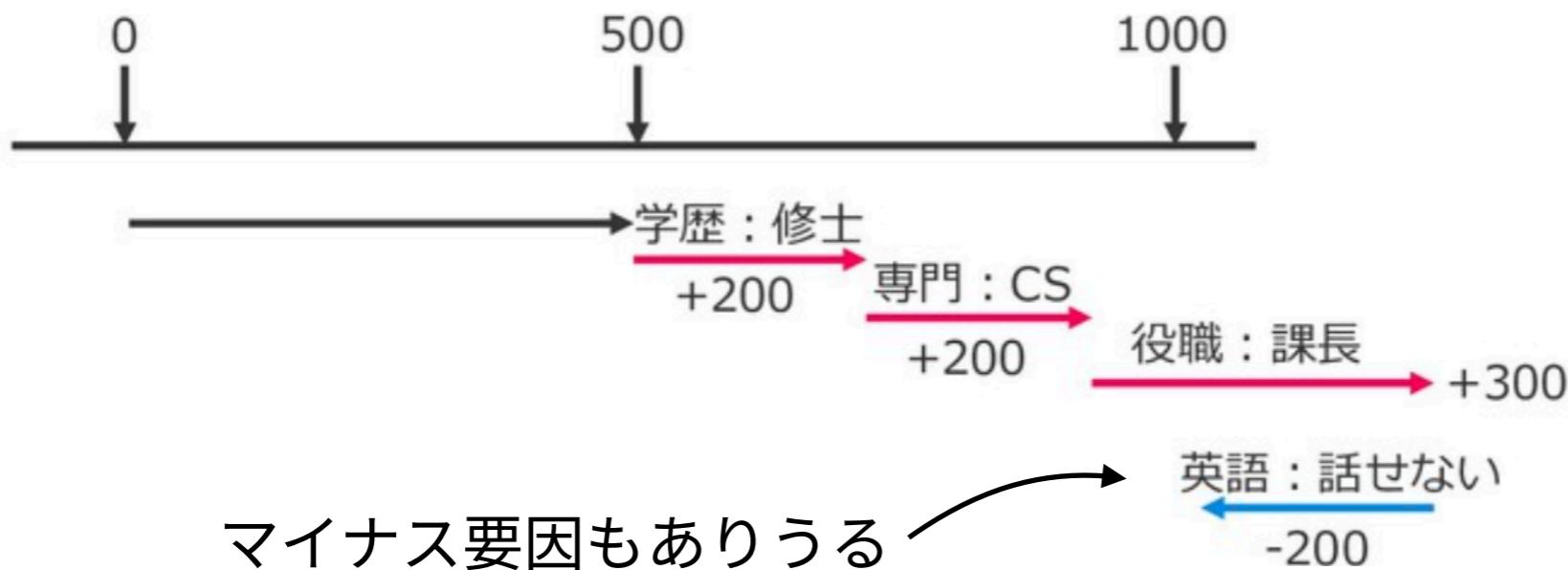


機械学習モデルがなぜその予測をしたかの要因分析

SHAP (SHapley Additive exPlanations)

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

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



機械学習モデルがなぜその予測をしたかの要因分析

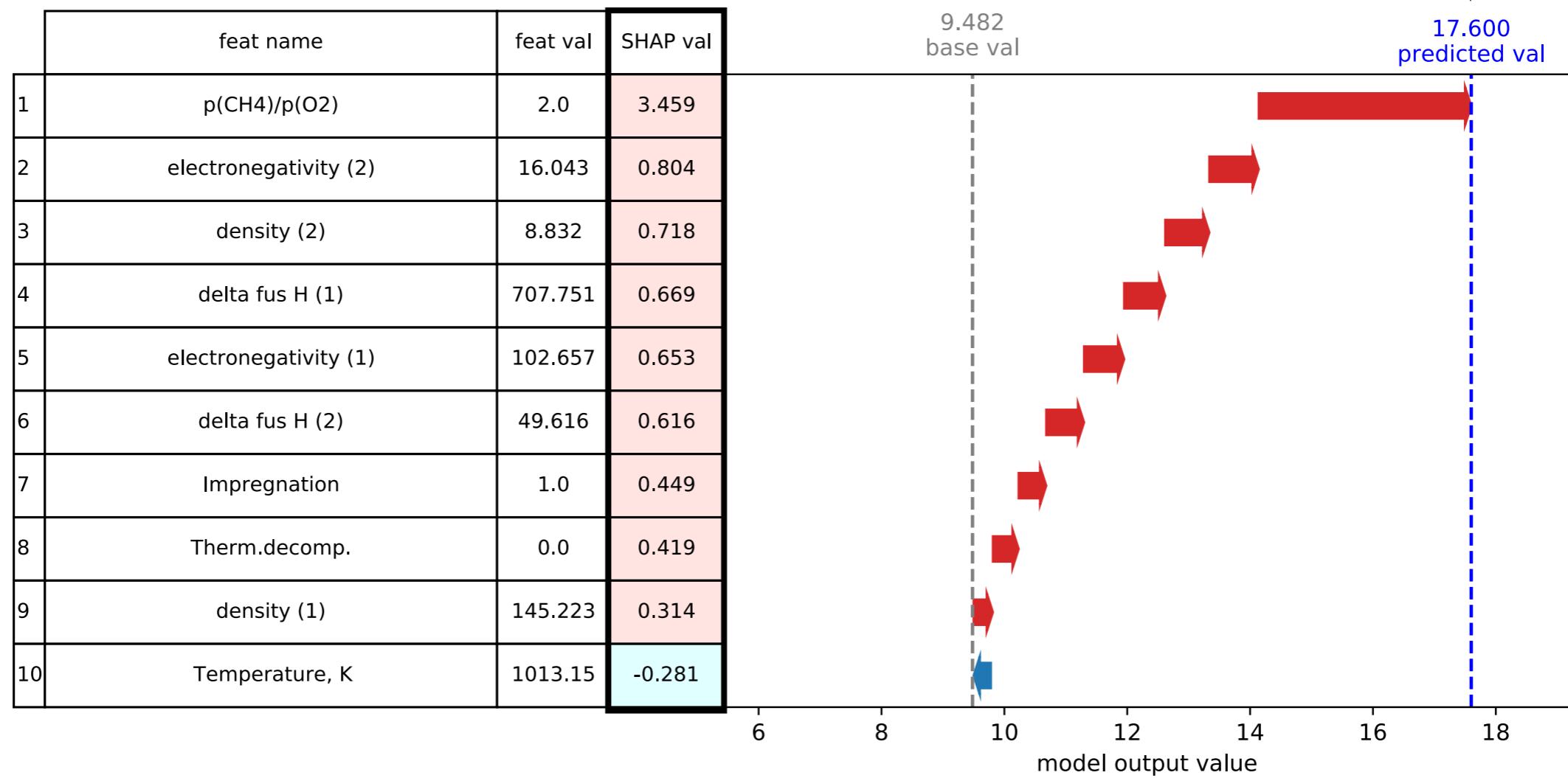
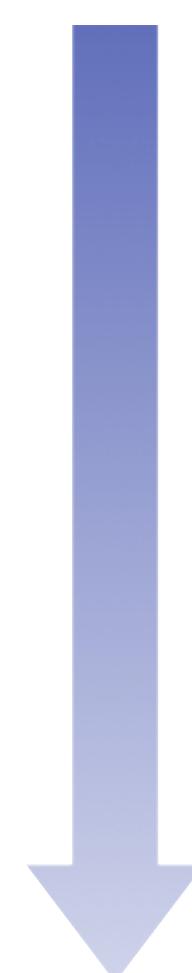
多人数の協力ゲームで得た報酬を各プレイヤへ公平に分配する
ゲーム理論の問題とみなすことで各々の特徴量の寄与度を算出

|SHAP値|

の降順

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

SHAP値の和 = データセット平均からの増加分



TreeExplainer: 決定木アンサンブル用のSHAP

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

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

ARTICLES

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

nature
machine intelligence

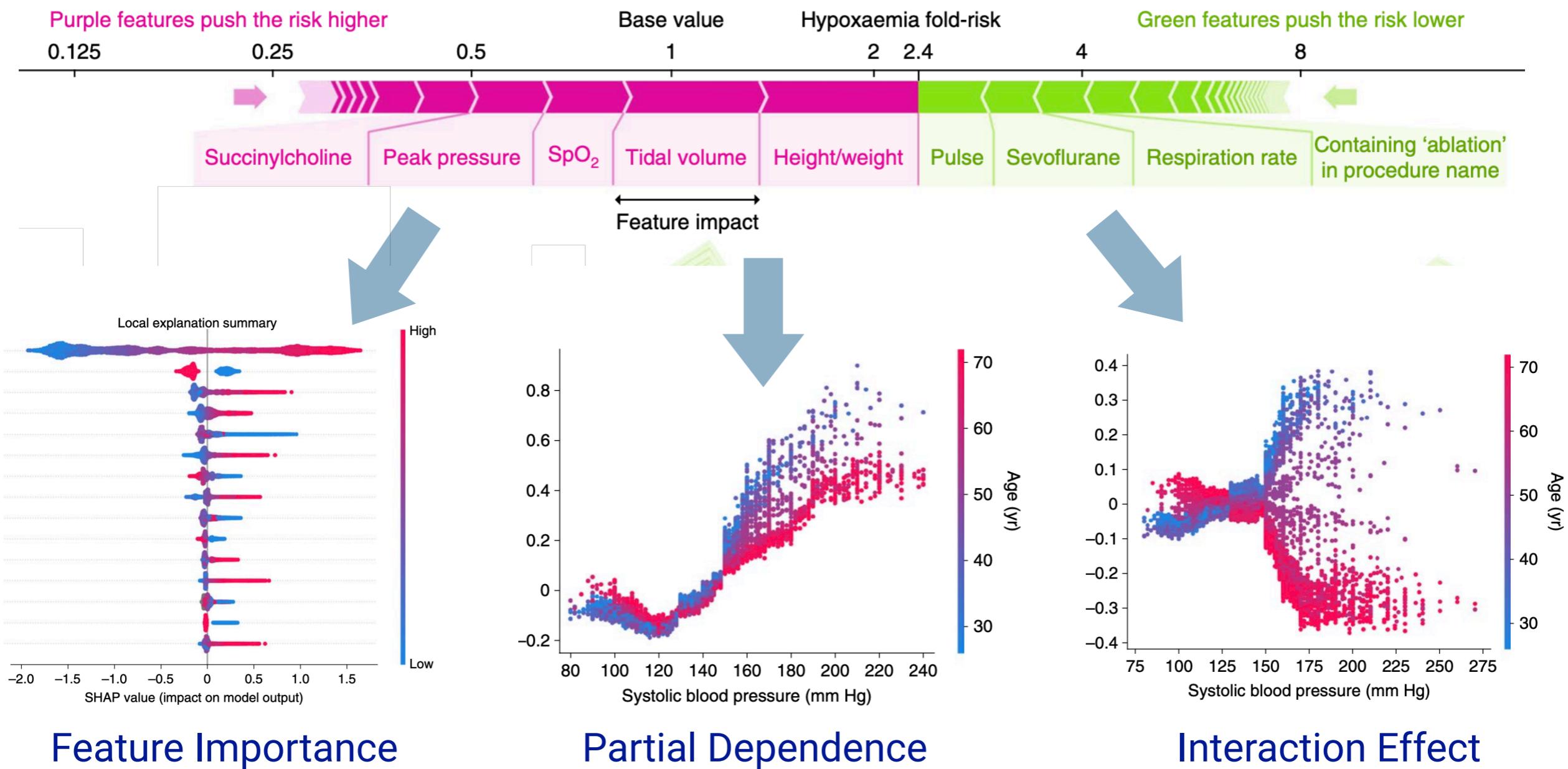
NATURE MACHINE INTELLIGENCE | VOL 2 | JANUARY 2020 | 56-67

From local explanations to global understanding with explainable AI for trees

Scott M. Lundberg^{1,2}, Gabriel Erion^{2,3}, Hugh Chen², Alex DeGrave^{2,3}, Jordan M. Prutkin⁴, Bala Nair^{5,6}, Ronit Katz⁷, Jonathan Himmelfarb⁷, Nisha Bansal⁷ and Su-In Lee^{2*}

SHAPによる学習済みモデルからの要因分析

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



このスライドpdfはここにあります

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

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