

IBISML 2022.12.22 @ 京大

機械学習と機械発見

自然科学研究におけるデータ利活用の再考

瀧川一学

ichigaku.takigawa@riken.jp

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

自己紹介：瀧川一学(たきがわいちがく)

機械学習の研究者であると同時に機械学習のユーザ

関心：離散構造を伴う機械学習 + 自然科学における機械発見

来歴

- 北大 工学研究科 博士(工学)
 - 京大 化学研究所/薬学研究科 助教 (7年)
 - 北大 情報科学研究科 准教授 (7年)
 - JST さきがけ (3.5年)
- 劣決定情報源分離の理論解析
バイオインフォマティクス・創薬化学
離散構造を伴う機械学習 (大規模知識処理)
材料インフォマティクス

現職 (クロスマポイントメント)

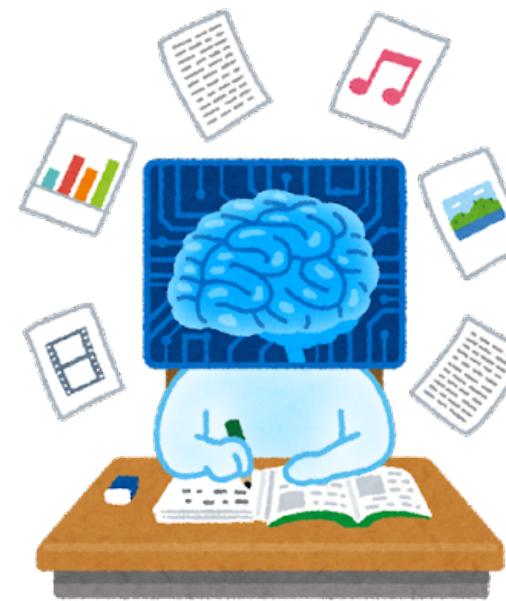
- 理研 革新知能統合研究センター 研究員 (4年～)
 - 北大 化学反応創成研究拠点 特任准教 (4年～)
- 機械学習と幹細胞生物学
機械学習と化学(計算化学・実験化学)

職場は理研AIPの一部チームが拠点を置く京阪奈ATR (住所は京都府だけどほぼ奈良県)

Materials Informatics

生命科学(Bioinformatics)には関わってきたが材料科学は全く「？」であった私は
「Materials Informatics」なる謎の分野に最初は楽観的イメージを持っていた…

ステップ①



利用できる色々なデータ
を機械学習に教えこむ

ステップ②



機械学習(エーアイくん)が
並の専門家より賢くなる

ステップ③



機械学習が有望な材料を
どんどん提案してくれる

今日のたった3つの話

お題「材料科学における機械学習」に沿って、この幻想が打ち砕かれてゆく過程で、機械学習研究者として得た教訓と最新知見をもとに、次の3つの主張をしてみたいと思います…

1. 「機械学習」と「材料科学」はゴールが根本的に食い違っていて、機械学習とは全く設定が異なる「**機械発見**」が求められている。
2. 仮説フリーの探索では、**決定木アンサンブルによる探索**がoff-the-shelfかつ非常に強力なベースラインになる。
3. それ以上を求めるなら、**仮説フリーではいられない。**

今日のたった3つの話

お題「材料科学における機械学習」に沿って、この幻想が打ち砕かれてゆく過程で、機械学習研究者として得た教訓と最新知見をもとに、次の3つの主張をしてみたいと思います…

1. 「機械学習」と「材料科学」はゴールが根本的に食い違っていて、機械学習とは全く設定が異なる「機械発見」が求められている。
2. 仮説フリーの探索では、決定木アンサンブルによる探索がoff-the-shelfかつ非常に強力なベースラインになる。
3. それ以上を求めるなら、仮説フリーではいられない。

今日の主張①

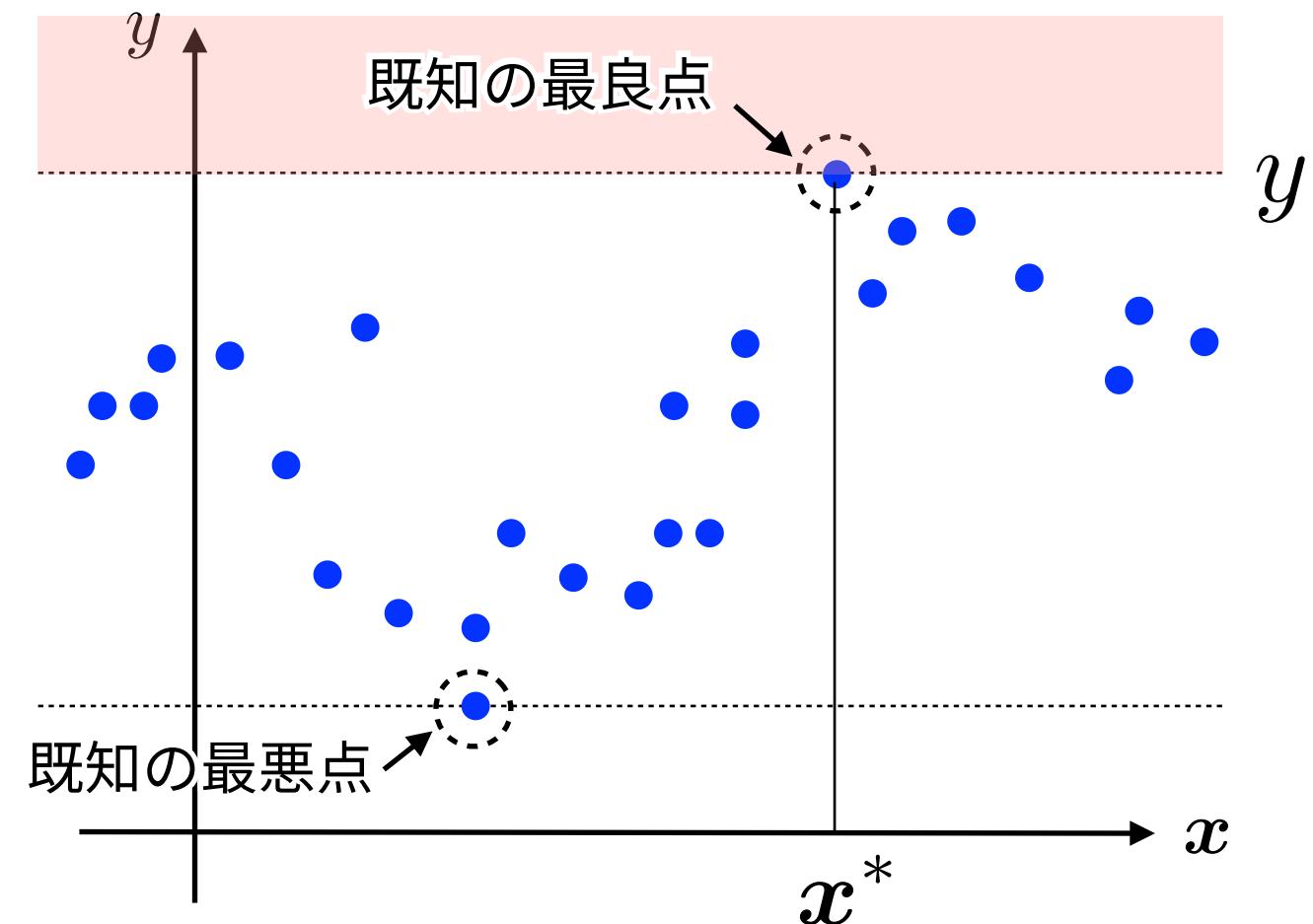
「機械学習」と「材料科学」はゴールが根本的に食い違っていて、機械学習とは全く設定が異なる「機械発見」が求められている。

- 機械学習は「訓練時に見せた見本例」と同じ統計的性質を仮定して予測を行う。
- 材料科学では基本的に「今あるどの材料よりも良い材料」を探したい。
 - 冷静に考えて今使われている材料より「一つも良い点がない材料」には価値がない。
 - 統計学点視点に立てば、これは「外れ値(例外)が欲しい」と言っているに等しい。(今知られている材料の中で最も良い材料でもかなり例外的だがそれより外れ値)

機械学習の基本設定：「期待(≈平均)」 誤差の最小化

今持っているデータから「外れ値」を探す

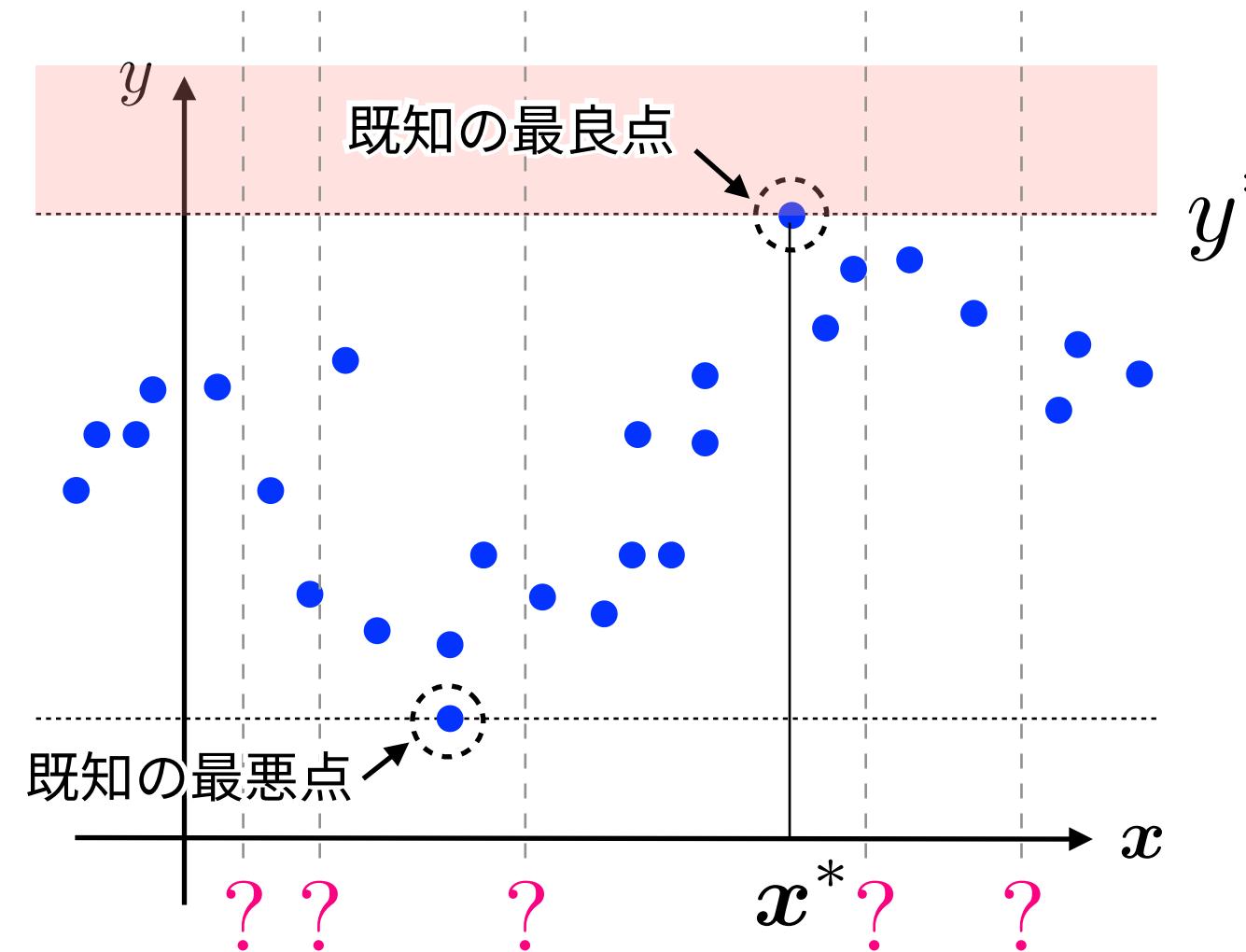
既知の最良点 x^* を超える x はどこ？



機械学習の基本設定：「期待(≈平均)」 誤差の最小化

今持っているデータから「外れ値」を探す

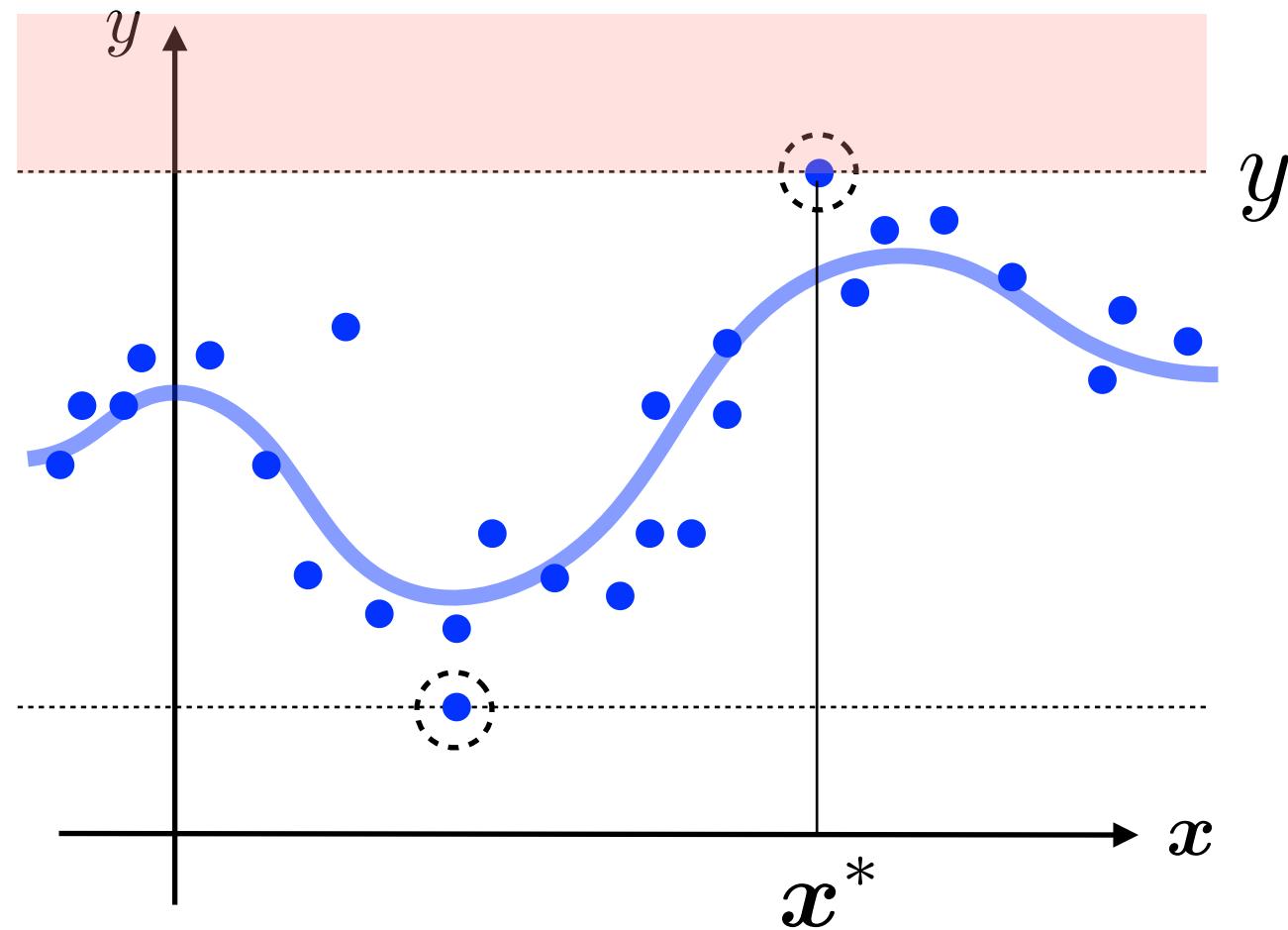
既知の最良点 x^* を超える x はどこ？



機械学習の基本設定：「期待(≈平均)」 誤差の最小化

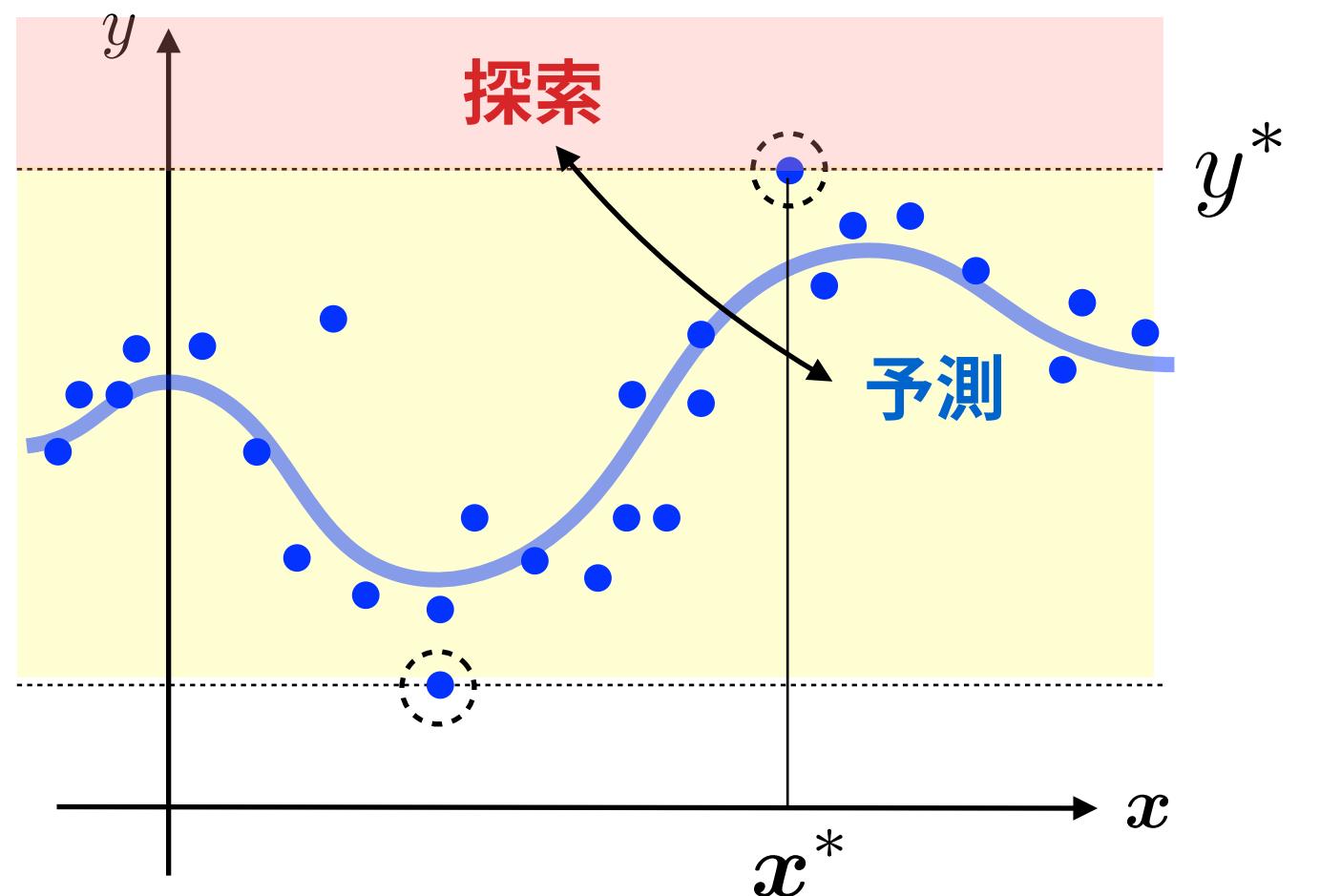
今持っているデータから「外れ値」を探す
既知の最良点 x^* を超える x はどこ？

今持っているデータを機械学習してみる
→ 予測 $\hat{y} = f(x)$ を与える関数 f を得る。



機械学習の基本設定：「期待(≈平均)」 誤差の最小化

今持っているデータから「外れ値」を探す
既知の最良点 x^* を超える x はどこ？



今持っているデータを機械学習してみる
→ 予測 $\hat{y} = \textcolor{blue}{f}(x)$ を与える関数 f を得る。

まともな機械学習手法では、どこかの x で

$$\hat{y} = \textcolor{blue}{f}(x) > y^*$$

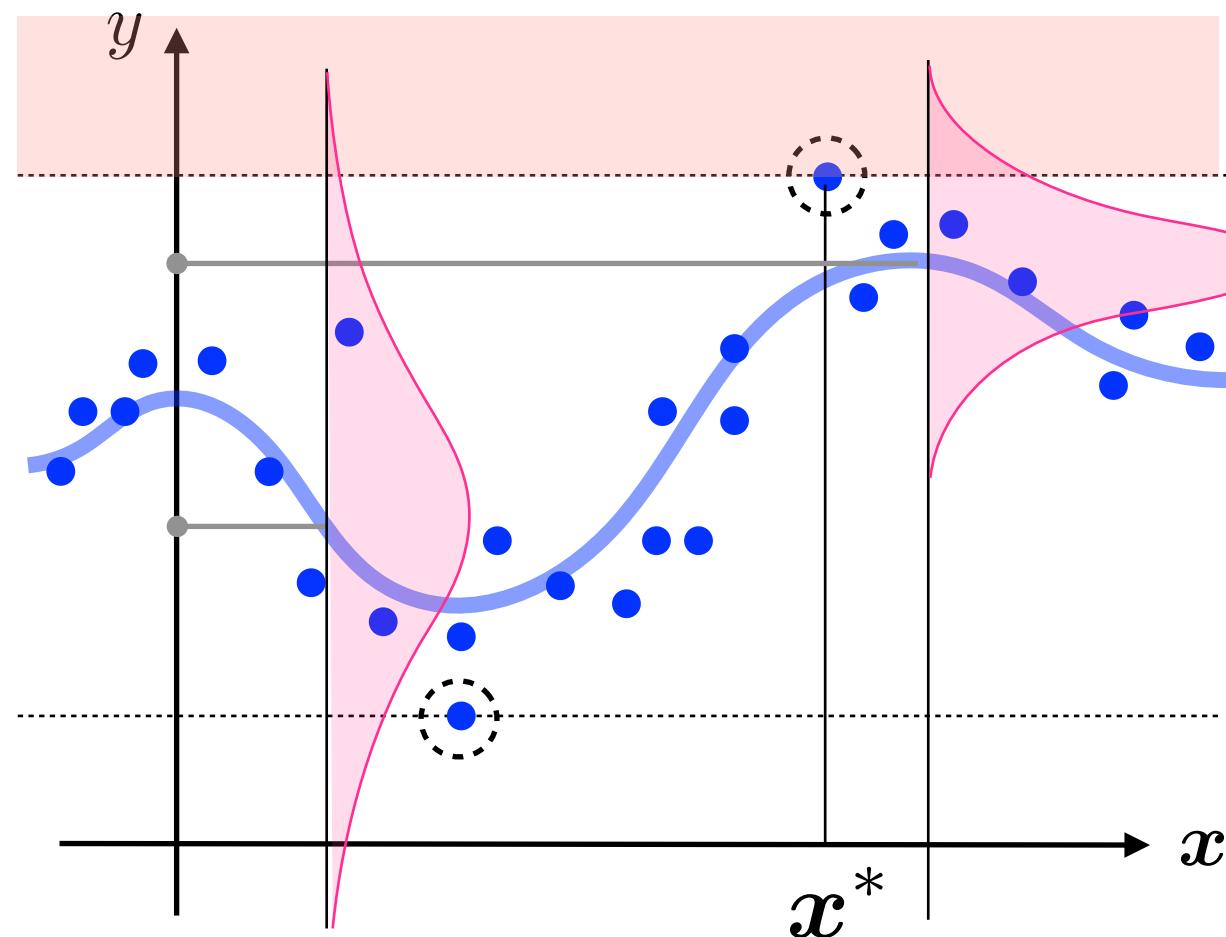
となることは基本的に期待できない。

誤差の期待値を最小化 = 見本例の真ん中を通る！

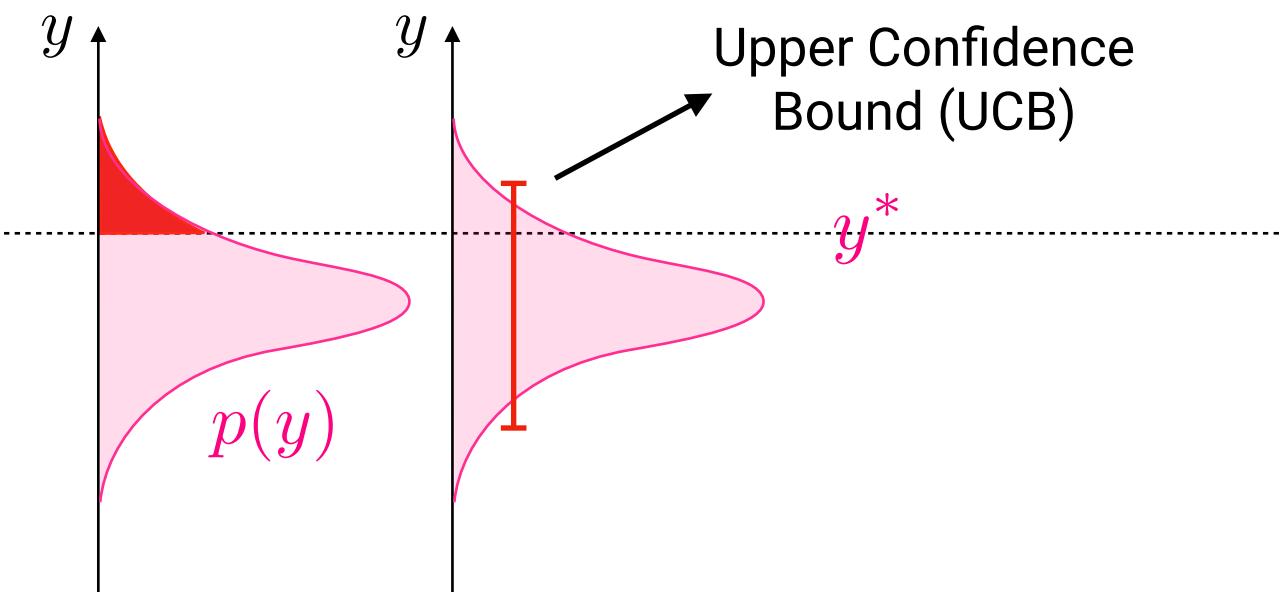
$$f = \arg \min_{f \in \mathcal{F}} \mathbb{E} [\text{error}(y, f(x))]$$

不確実性の定量化(Uncertainty Quantification, UQ)

今持っているデータから「外れ値」を探す
既知の最良点 x^* を超える x はどこ？



予測値そのものは探索指標にならないので
「不確実性の定量化(UQ)」 が根本的に重要



Probability of improvement (PI)

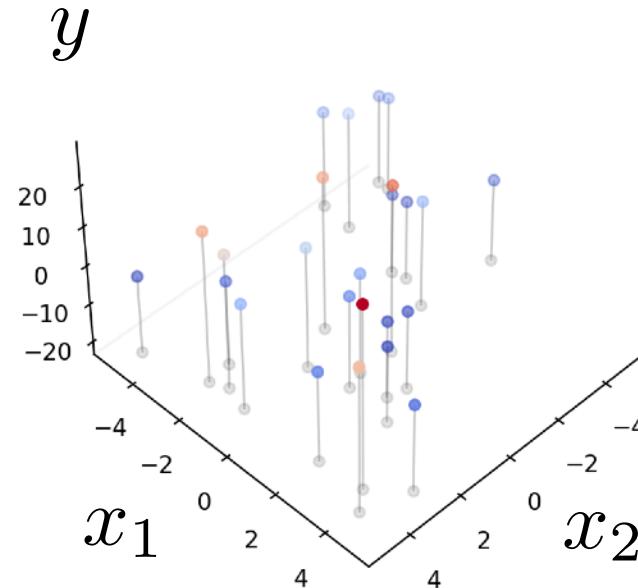
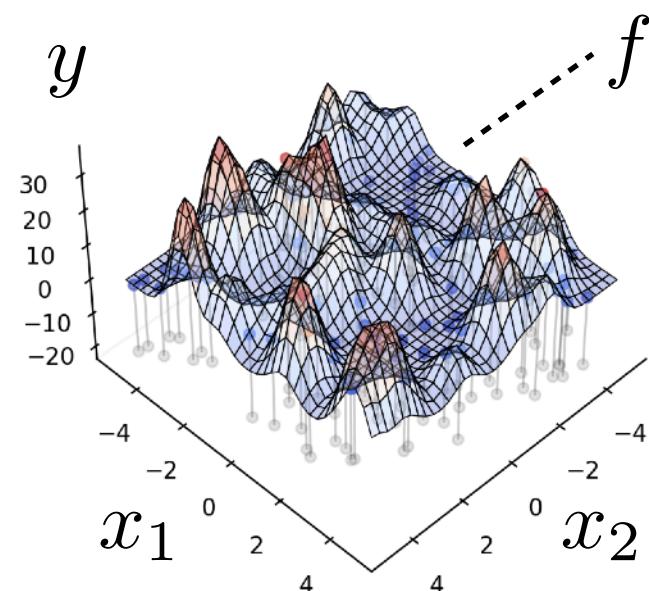
$$P(y > y^*) = \int_{y^*}^{\infty} p(y) dy$$

Expected Improvement (EI)

$$\mathbb{E}[y | y > y^*] = \int_{y^*}^{\infty} y \cdot p(y) dy$$

材料科学と機械学習での「テストデータ」の違い

さらに機械学習の根幹である「**テストデータ**」の概念がそもそも絶望的に異なる！

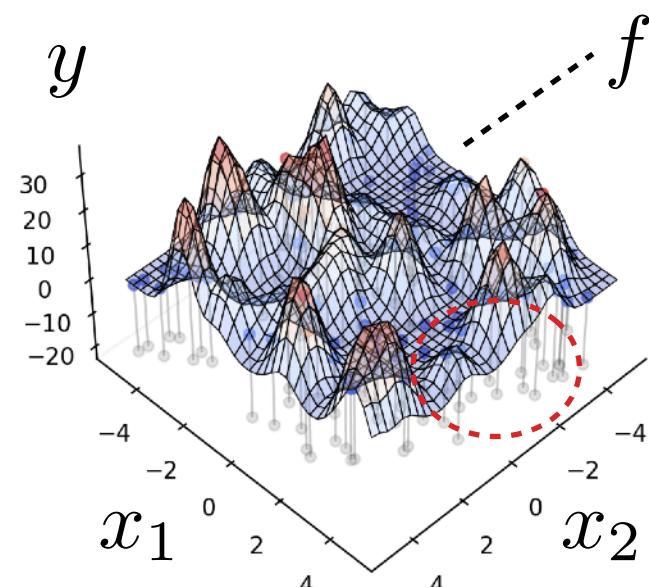


$$x_1 \rightarrow \boxed{f} \rightarrow y$$

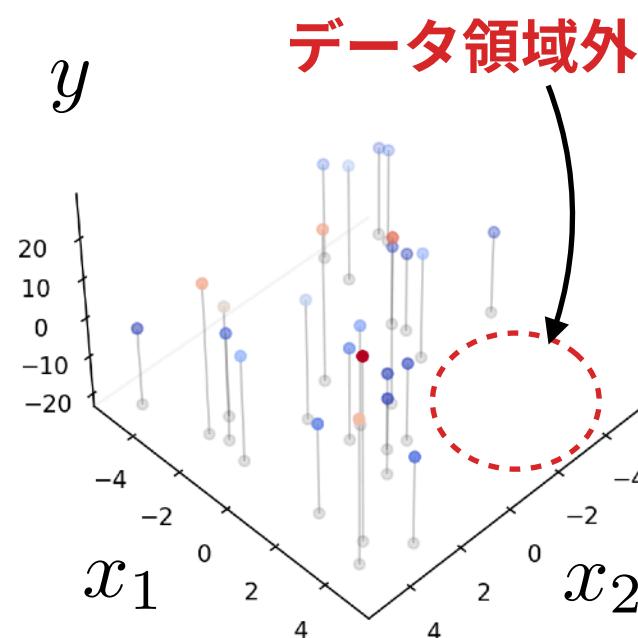
得られているデータ
(訓練データ)

材料科学と機械学習での「テストデータ」の違い

さらに機械学習の根幹である「**テストデータ**」の概念がそもそも絶望的に異なる！



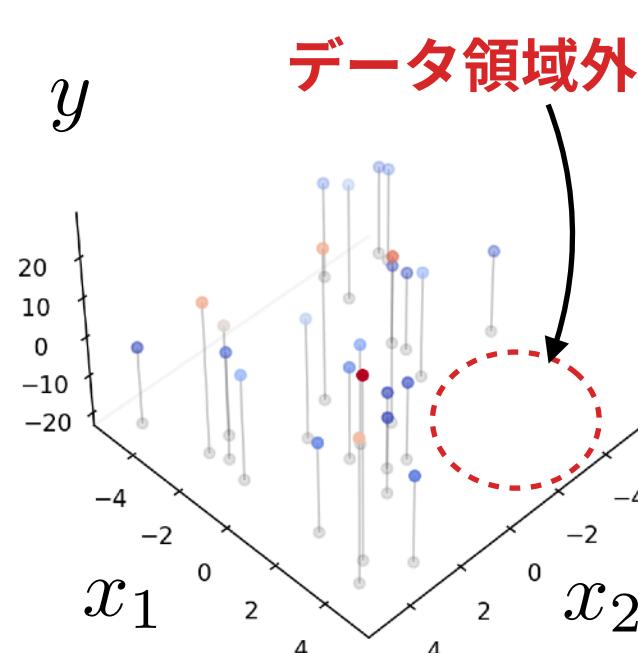
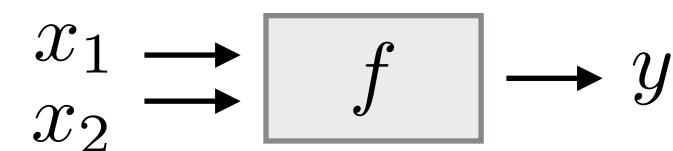
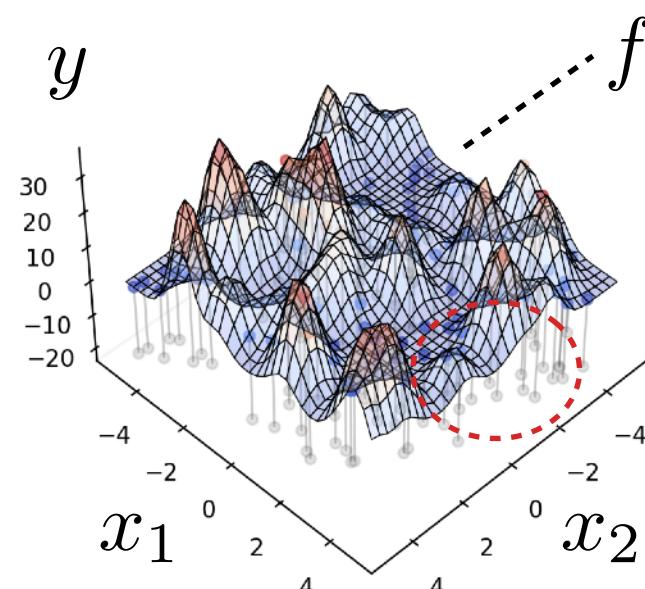
$$x_1 \rightarrow \boxed{f} \rightarrow y$$
$$x_2 \rightarrow$$



得られているデータ
(訓練データ)

材料科学と機械学習での「テストデータ」の違い

さらに機械学習の根幹である「**テストデータ**」の概念がそもそも絶望的に異なる！



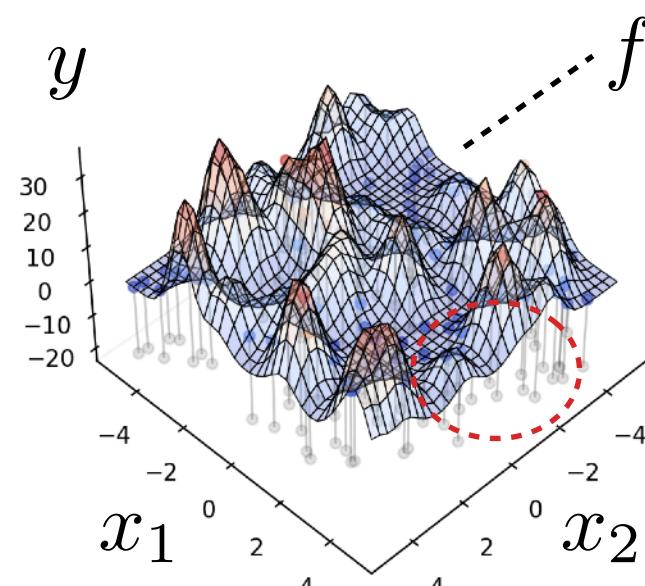
得られているデータ
(訓練データ)

データ領域外 **機械学習** **データ領域外**は諦める(小確率だし)

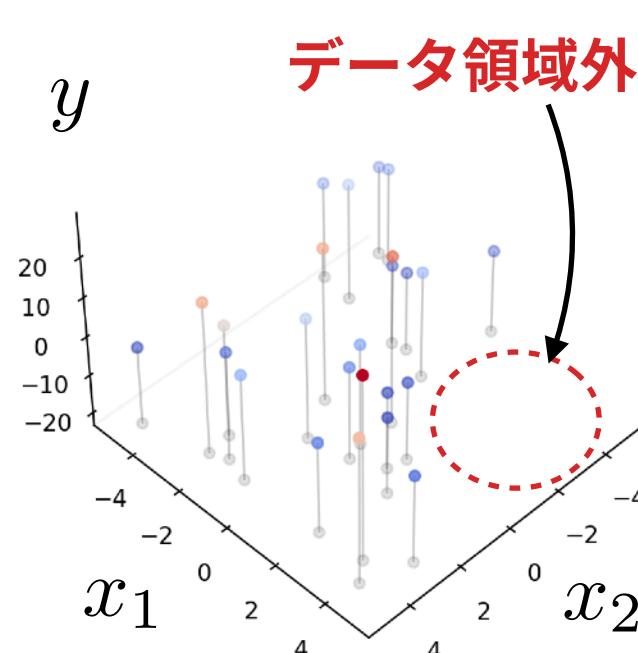
- 得られているデータが分布している領域に (=得られているデータとほぼ同じ確率分布で) 来た点 (x_1, x_2) での y を予測する

材料科学と機械学習での「テストデータ」の違い

さらに機械学習の根幹である「**テストデータ**」の概念がそもそも絶望的に異なる！



$$\begin{matrix} x_1 \\ x_2 \end{matrix} \rightarrow \boxed{f} \rightarrow y$$



得られているデータ
(訓練データ)

機械学習 データ領域外は諦める(小確率だし)

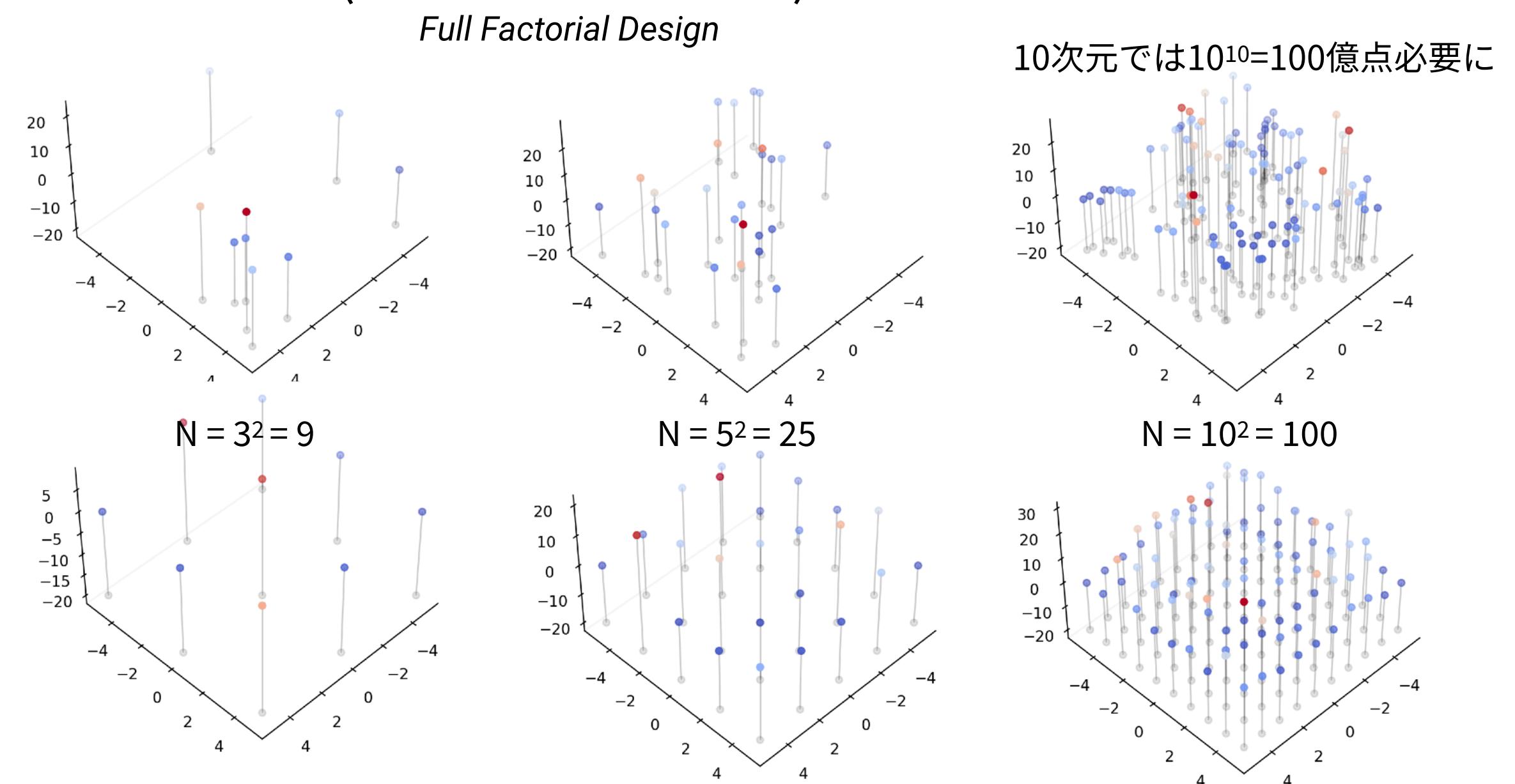
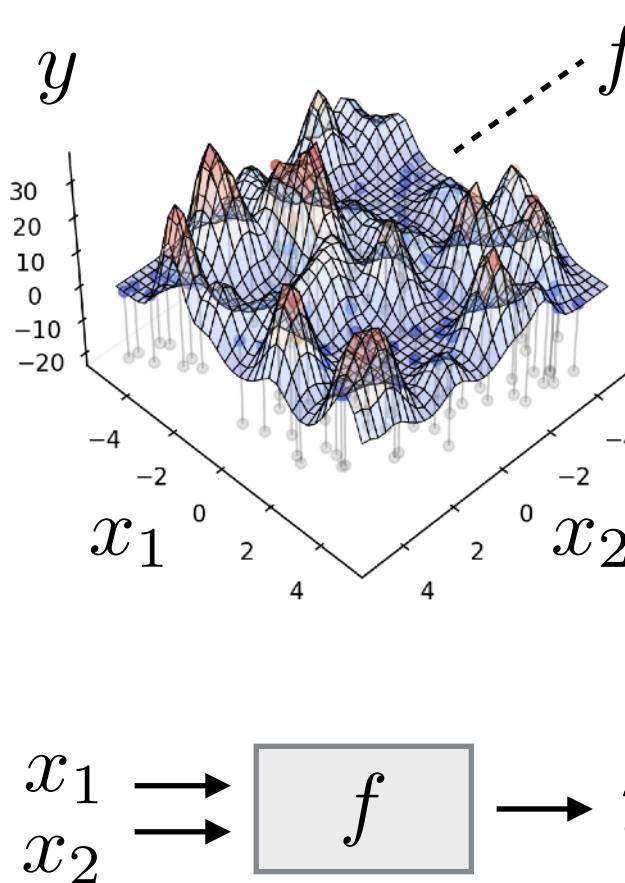
- 得られているデータが分布している領域に (=得られているデータとほぼ同じ確率分布で) 来た点 (x_1, x_2) での y を予測する

材料科学 むしろデータ領域外に関心？

- 点 (x_1, x_2) は「どこでも良いので(探索)」大きな y を与える点 (x_1, x_2) が知りたい

材料科学と機械学習での「訓練データ」の違い

探索が目的なら「訓練データ」も探索範囲を満遍なくカバーするよう取るべき。
→ ただし、高次元では全組合せ実施(完全実施要因計画)は現実的には実施不可能



材料科学と機械学習での「訓練データ」の違い

実験計画法におけるFisherの三原則

1. 反復 (Replication)

自然現象を対象とする場合、同条件で複数回の実験が必要 (系統誤差と偶然誤差の判別)

2. 局所管理 (Local Control)

考慮している因子以外の背景因子はできるだけ均一になるように実験を管理する

3. 無作為化 (Randomization) → Randomized Controlled Trial (Fisherの奥義)

データバイアスや潜在的交絡因子からのData Leakageによる疑似相関で予測が当たってしまう事故を防ぐため、制御可能な因子はすべてランダムに割り付けを行う

これを全て満たす訓練データを得るのは(自動実験系以外では)現実的に難しい…

「機械発見」への道

予測技術である標準的な機械学習の前提とはかなり異なる 「機械発見」 の問題

ゴールは「学習」でも「予測」でもなく「発見」(or 「理解」?)

→ 機械学習はその部分要素技術の一つにすぎない

→ 訓練データの計画、テストデータの計画、UQ、高次元探索、帰納と演繹の融合、…

主張②

主張③

「機械発見」への道

予測技術である標準的な機械学習の前提とはかなり異なる 「機械発見」 の問題

ゴールは「学習」でも「予測」でもなく「発見」(or 「理解」?)

→ 機械学習はその部分要素技術の一つにすぎない

→ 訓練データの計画、テストデータの計画、UQ、高次元探索、帰納と演繹の融合、…

主張②

主張③

自然科学を試験台にすれば抽象論ではない 「機械発見」 の方法論を再び体系的に研究できる！

→ 「(科学的)発見」を合理化できるのか？は人工知能及び科学哲学の(永遠の?)未解決課題

Herbert A. Simon



有川 節夫



- Simon HA, *Machine Discovery*. (1997)
- Langley PW, Simon HA, Bradshaw G, Zytkow JM, *Scientific Discovery: Computational Explorations of the Creative Process* (1987).
- 有川節夫, 機械学習から機械発見へ. (1996)
- 森下真一・宮野悟編, *発見科学とデータマイニング*. (2001)

今日のたった3つの話

お題「材料科学における機械学習」に沿って、この幻想が打ち砕かれてゆく過程で、機械学習研究者として得た教訓と最新知見をもとに、次の3つの主張をしてみたいと思います…

1. 「機械学習」と「材料科学」はゴールが根本的に食い違っていて、機械学習とは全く設定が異なる「**機械発見**」が求められている。
2. 仮説フリーの探索では、**決定木アンサンブルによる探索**がoff-the-shelfかつ非常に強力なベースラインになる。
3. それ以上を求めるなら、**仮説フリーではいられない。**

今日の主張②

仮説フリーの探索では、**決定木アンサンブルによる探索**がoff-the-shelfかつ非常に強力なベースラインになる。

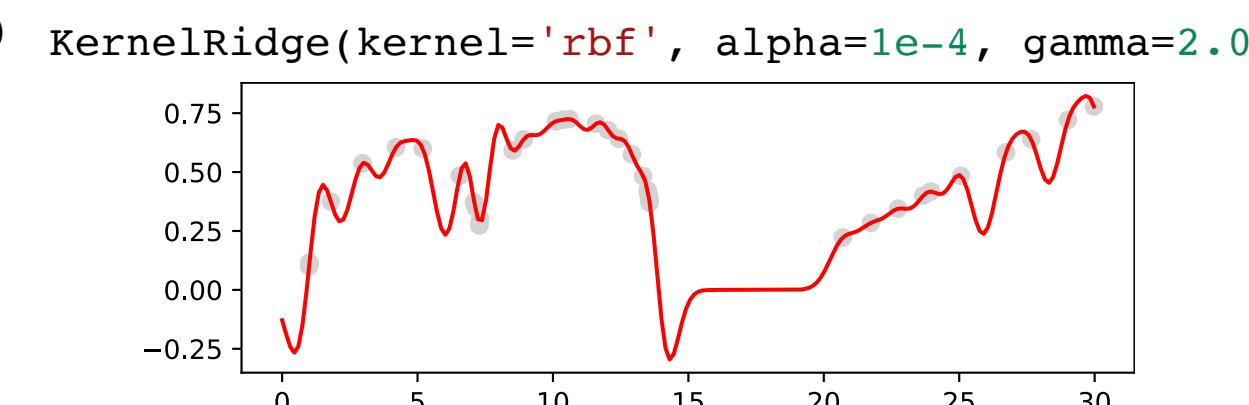
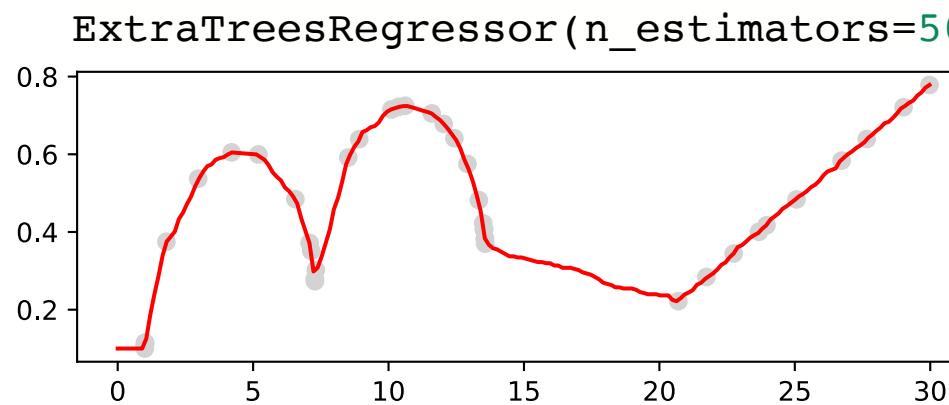
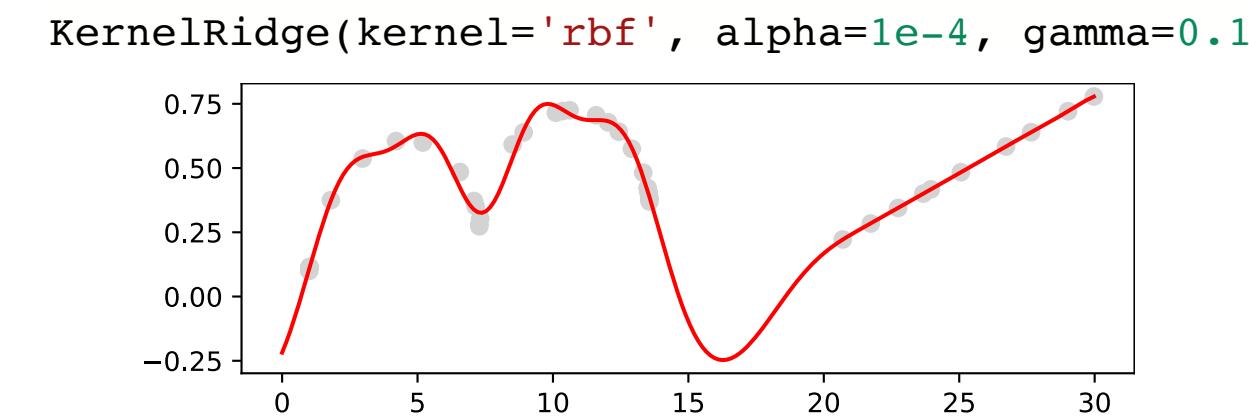
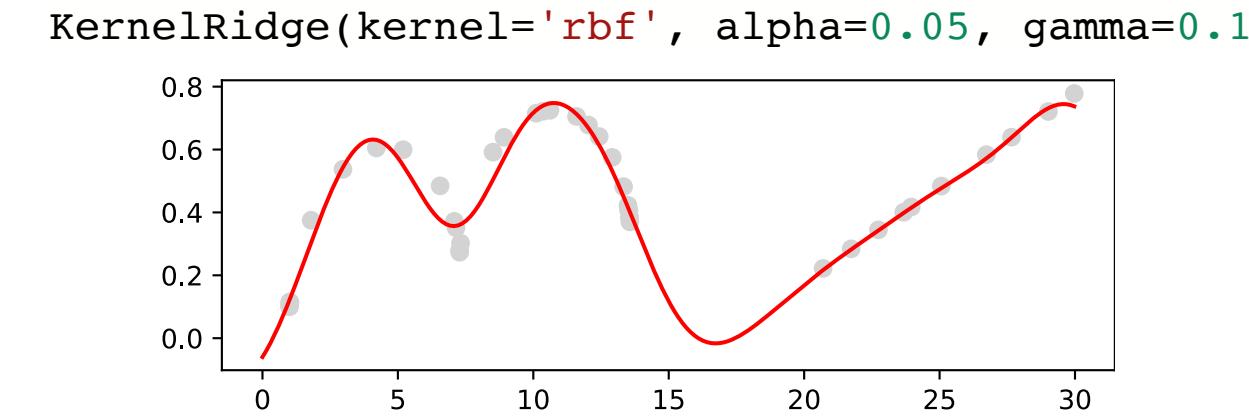
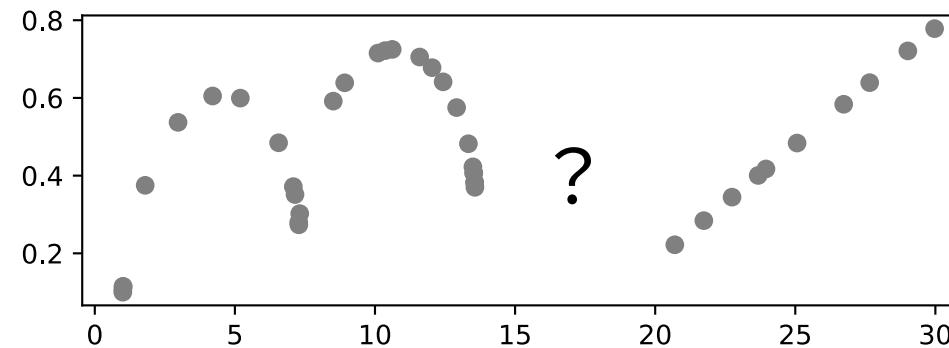
今日の主張②

仮説フリーの探索では、**決定木アンサンブルによる探索**がoff-the-shelfかつ非常に強力なベースラインになる。

- 探索が目的だとすると「高次元」の**任意の点**での予測が求められるが、有限の見本点は(“big data”だとしても)高次元空間ではスカスカで**「データ領域外」での挙動が支配的な影響を与える**。(ほとんどの検査点はデータ領域外になる)
- 決定木アンサンブル = データ依存型の領域分割上の**ヒストグラム近似**であり、データ領域外での挙動が**極めて保守的で安全**。また、どんな高次元でも各領域に最低限落ちるべきサンプル数の下限でセーフガードもできる。
- 近年注目の局所的領域分割による**interpolator**である。(他の代表的が深層学習)

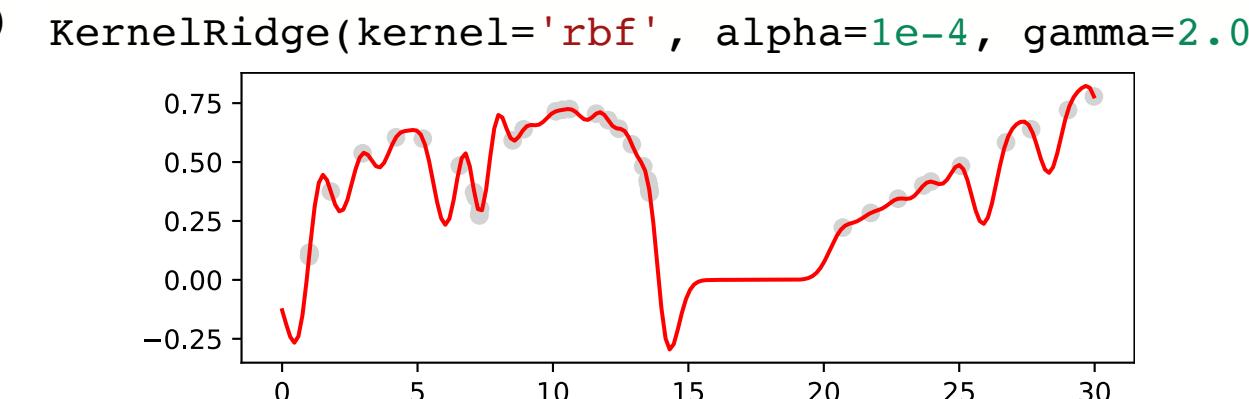
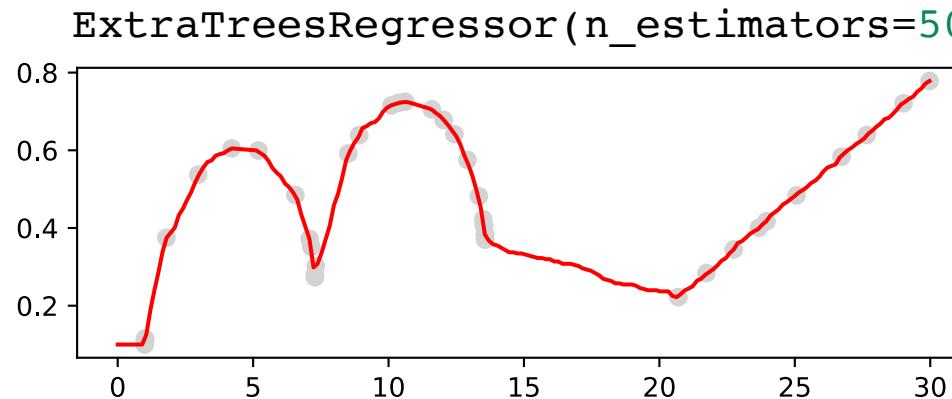
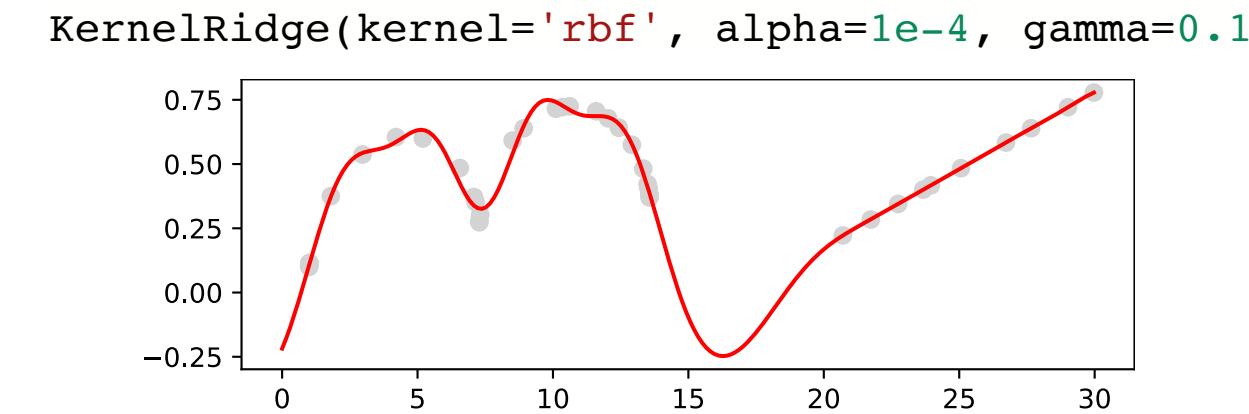
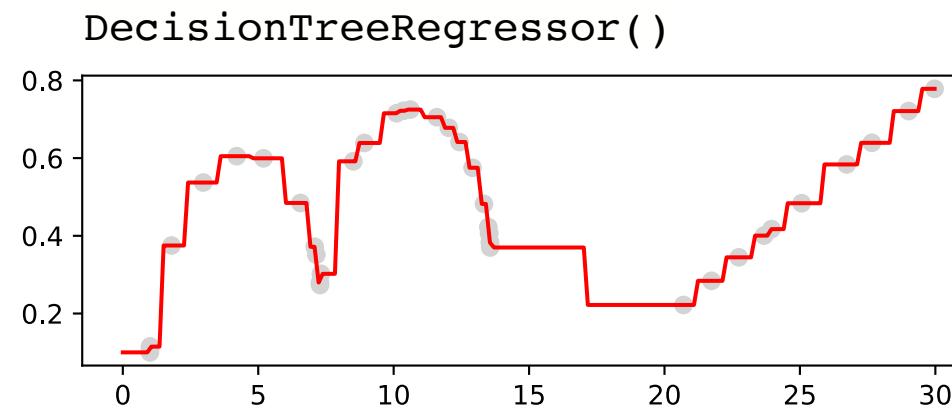
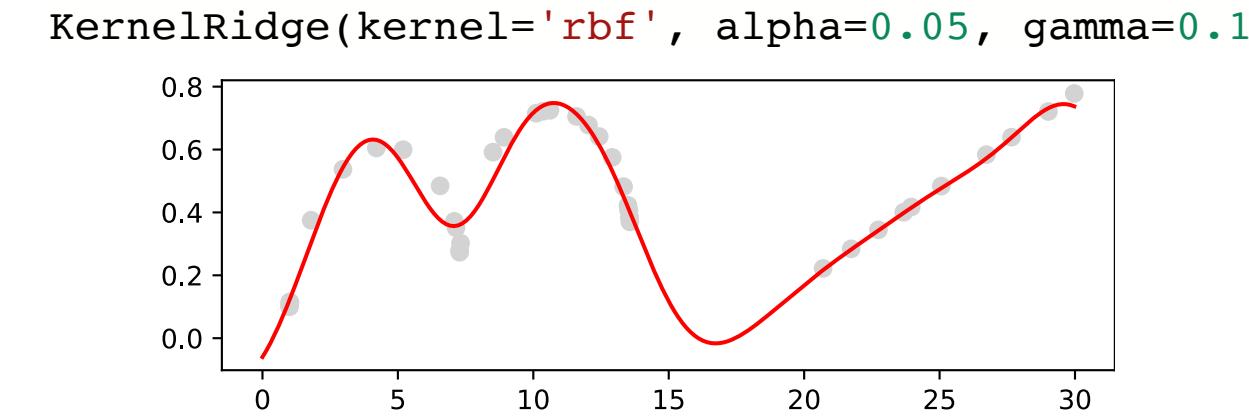
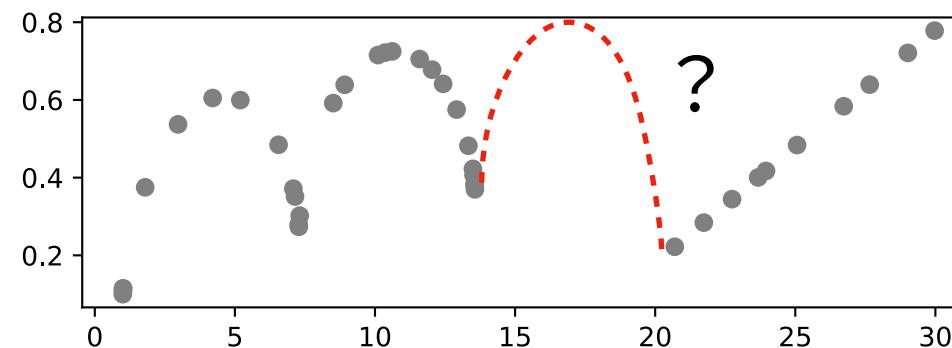
データ領域外の挙動？

結局「データがない」ので統計的手法はデータ領域外では積極的なことはできない…



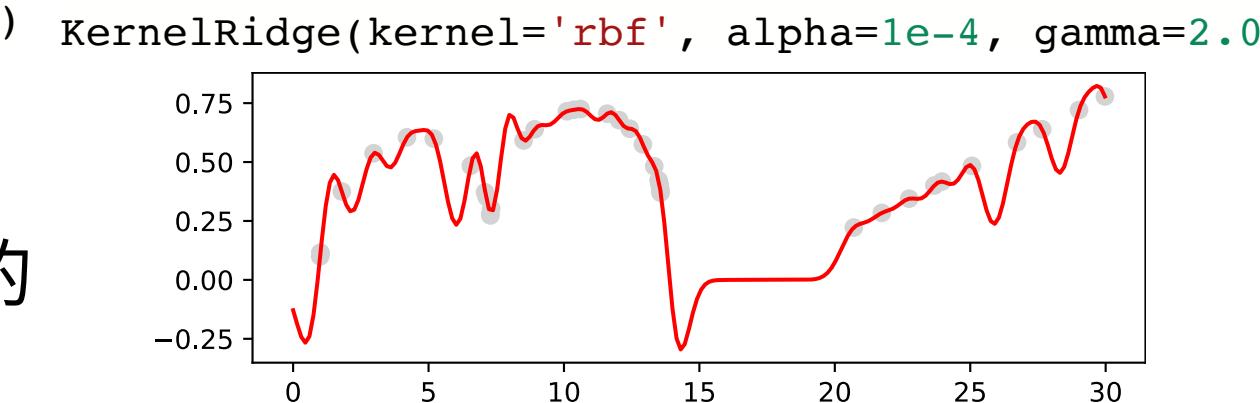
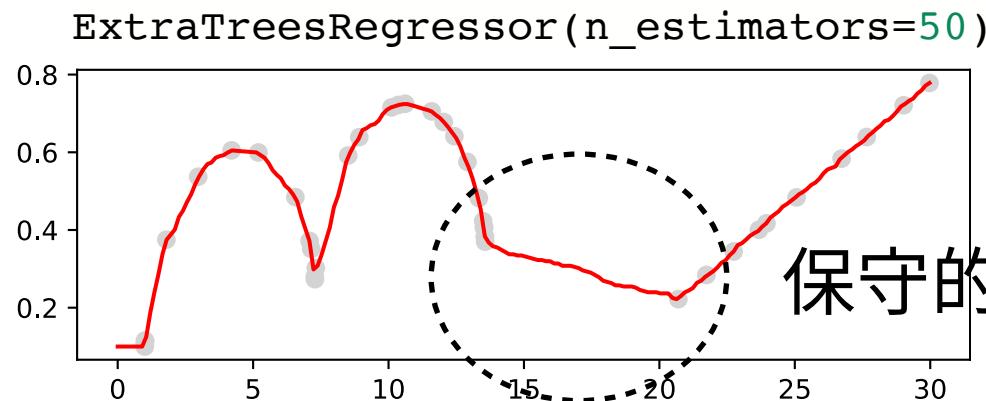
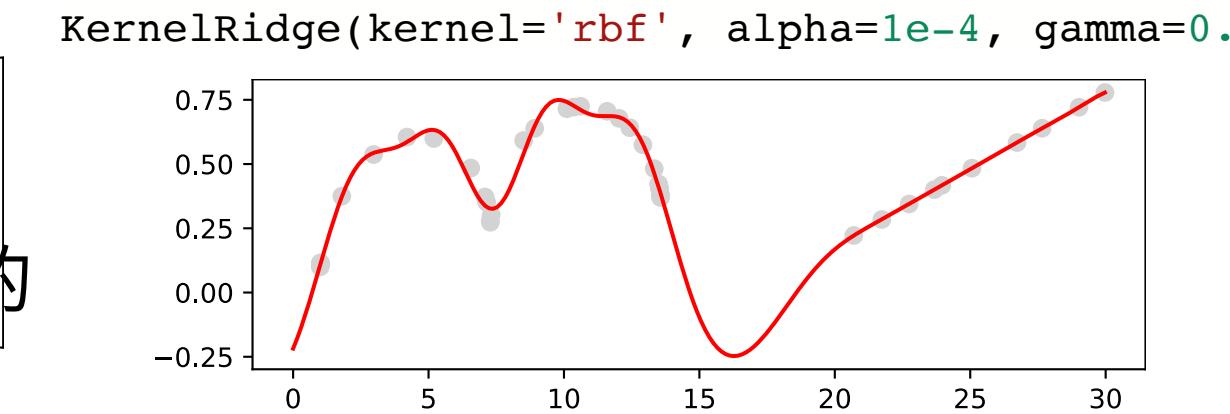
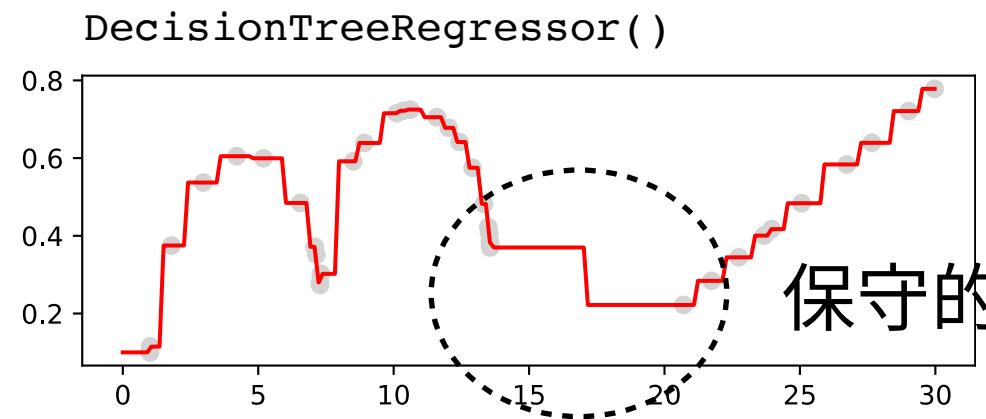
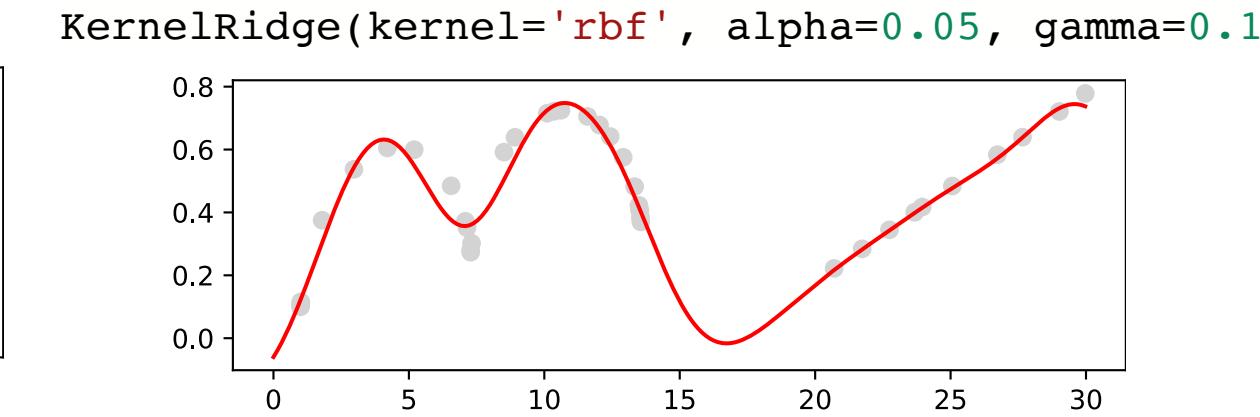
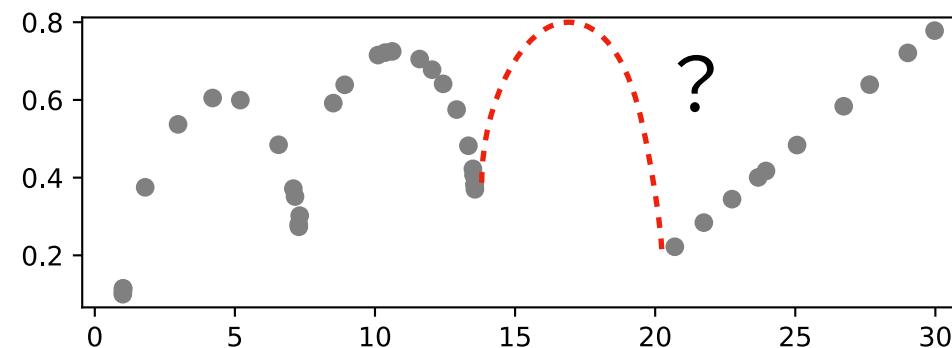
データ領域外の挙動？

結局「データがない」ので統計的手法はデータ領域外では積極的なことはできない…



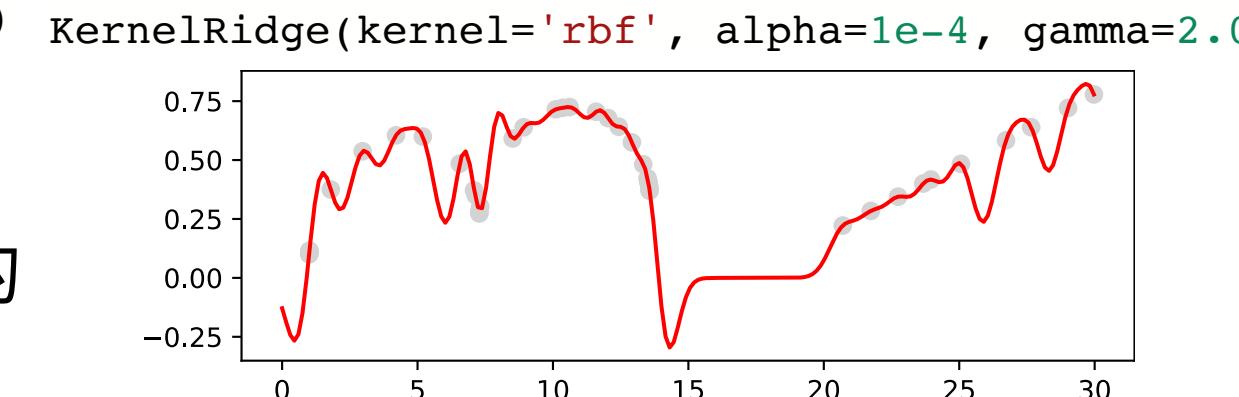
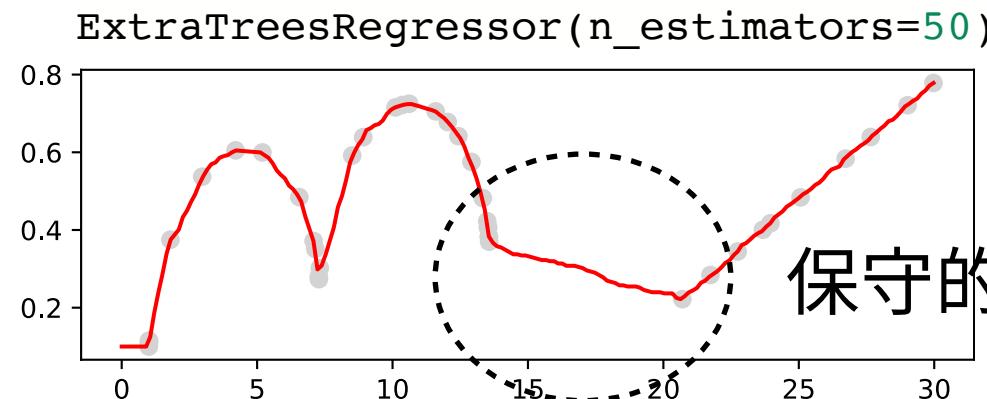
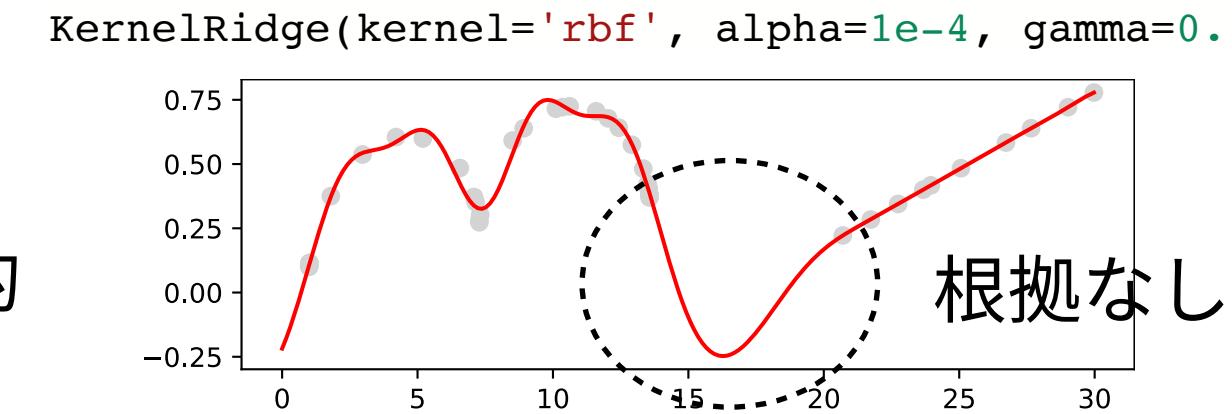
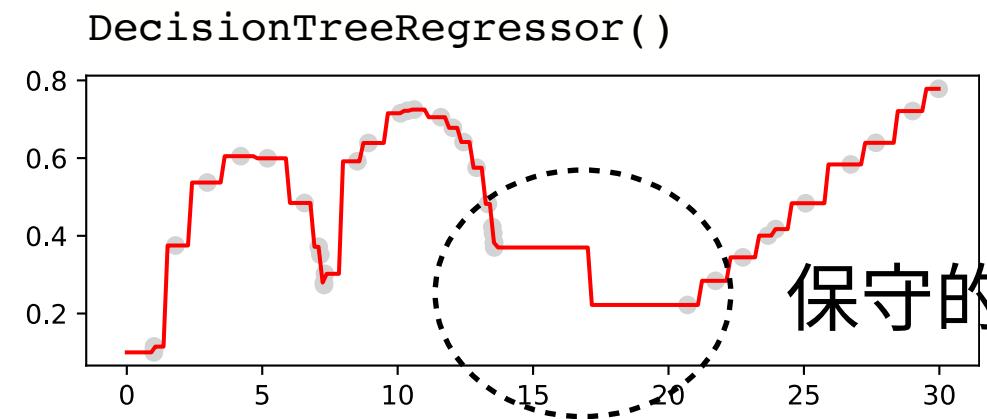
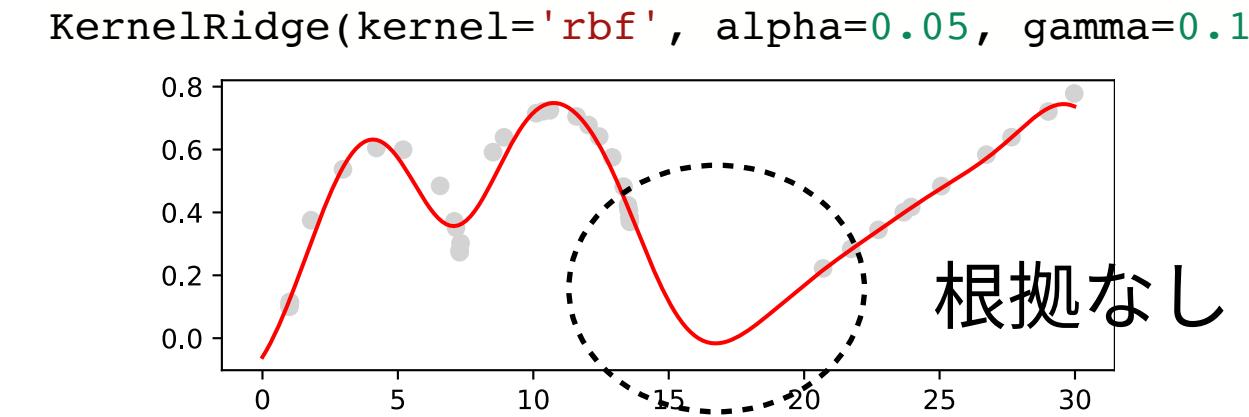
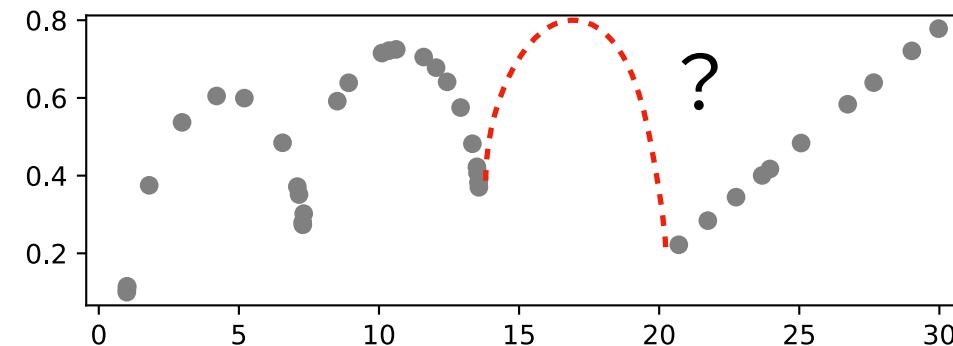
データ領域外の挙動？

結局「データがない」ので統計的手法はデータ領域外では積極的なことはできない…



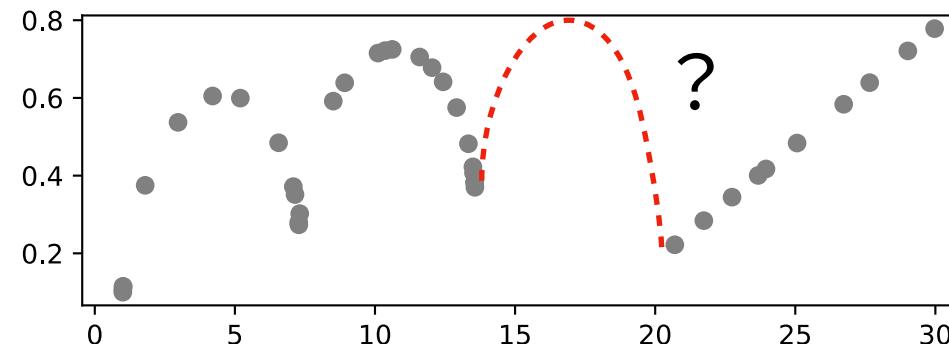
データ領域外の挙動？

結局「データがない」ので統計的手法はデータ領域外では積極的なことはできない…

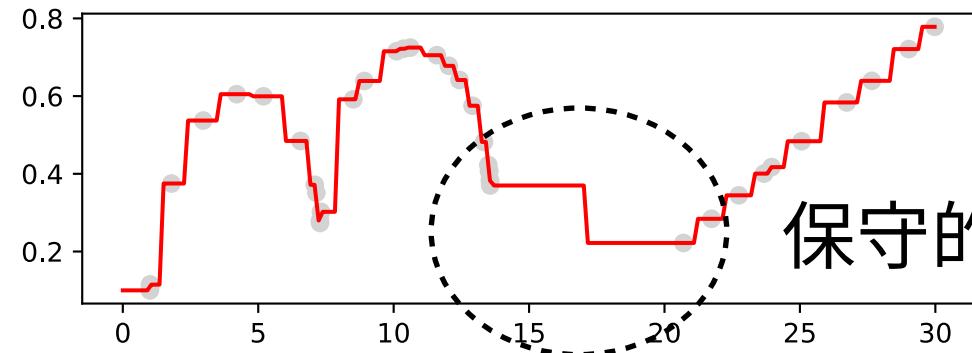


データ領域外の挙動？

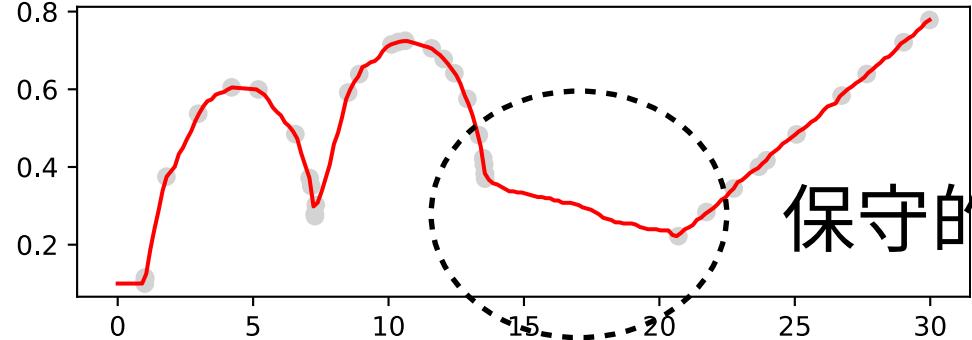
結局「データがない」ので統計的手法はデータ領域外では積極的なことはできない…



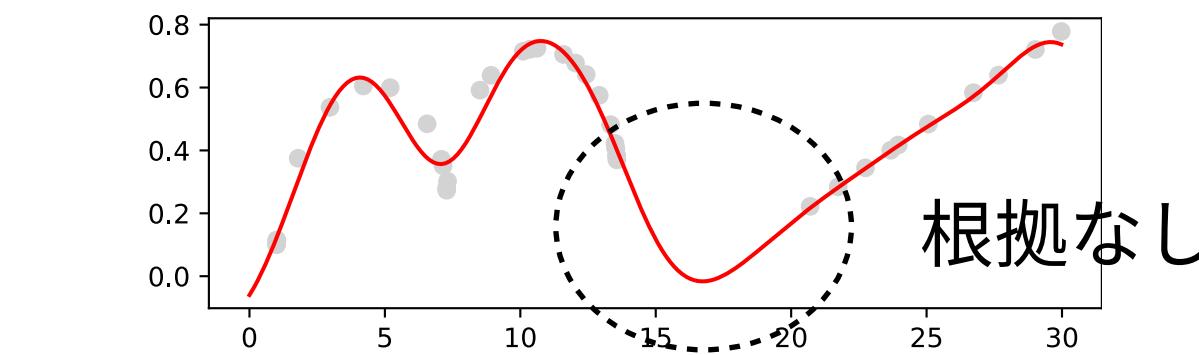
DecisionTreeRegressor()



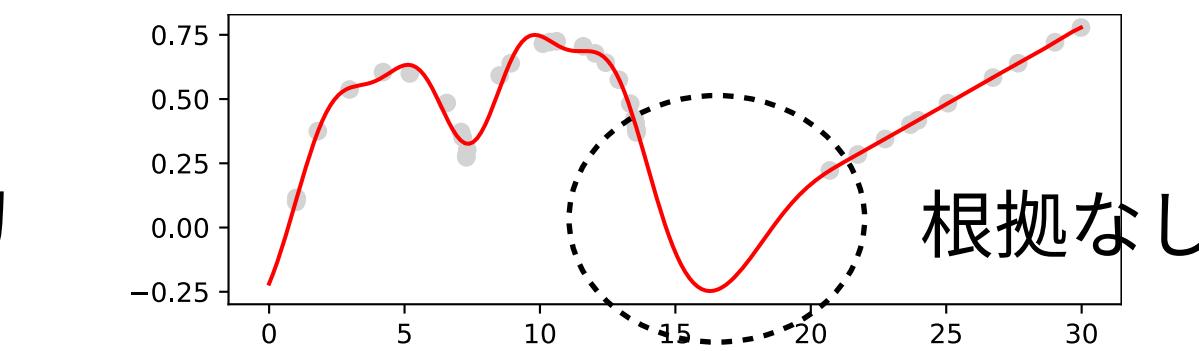
ExtraTreesRegressor(n_estimators=50)



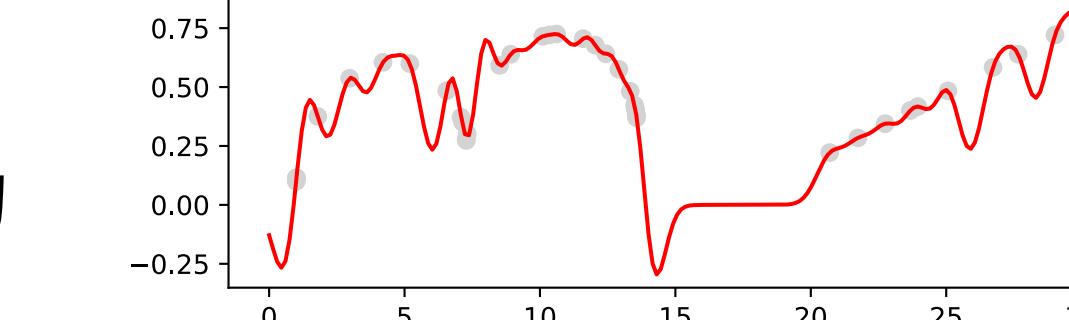
KernelRidge(kernel='rbf', alpha=0.05, gamma=0.1)



KernelRidge(kernel='rbf', alpha=1e-4, gamma=0.1)



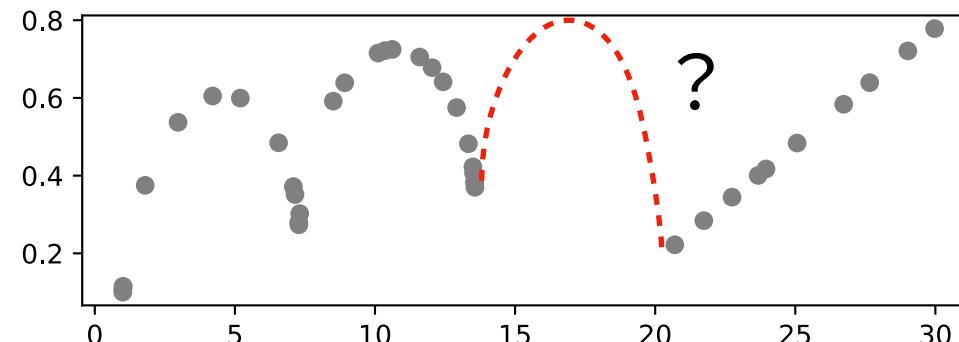
KernelRidge(kernel='rbf', alpha=1e-4, gamma=2.0)



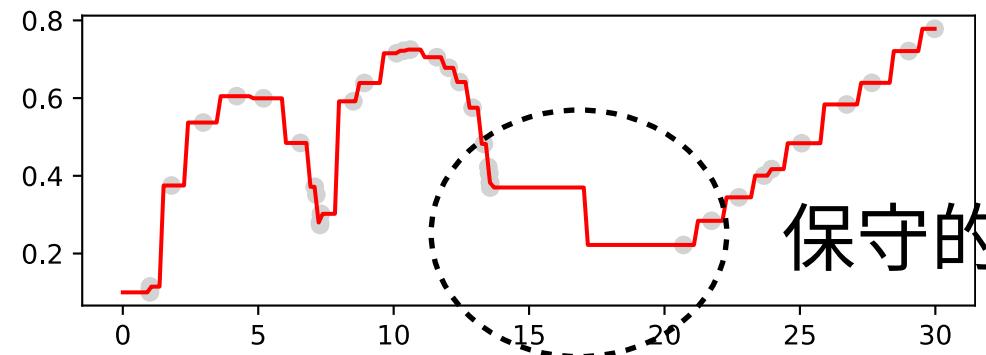
- この意図しない挙動はモデルの帰納バイアスに由来し検出が難しい（データがないので！）

データ領域外の挙動？

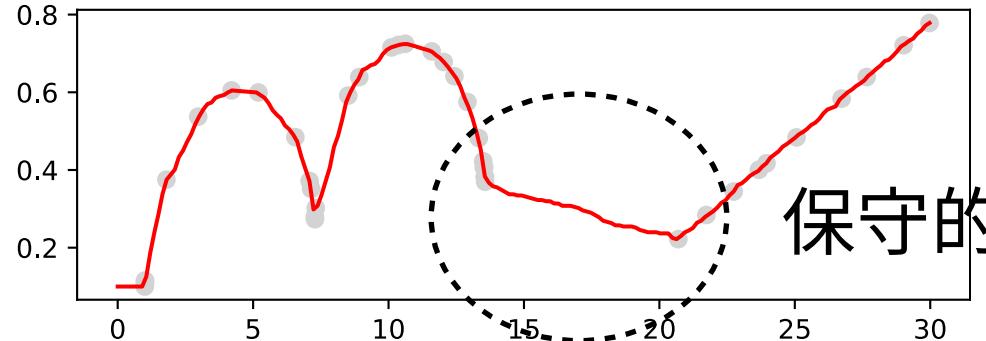
結局「データがない」ので統計的手法はデータ領域外では積極的なことはできない…



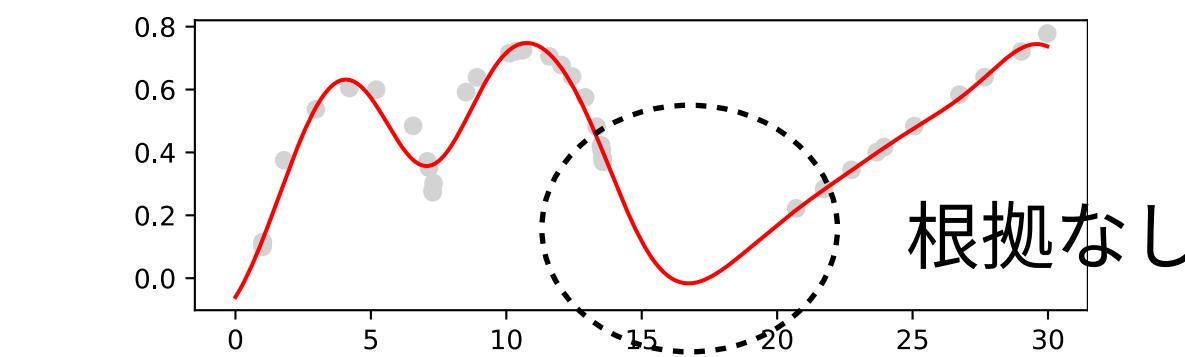
DecisionTreeRegressor()



ExtraTreesRegressor(n_estimators=50)

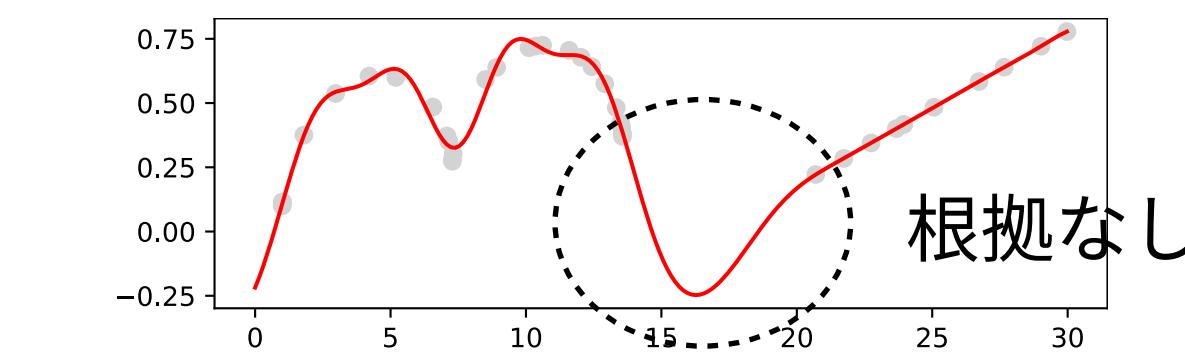


KernelRidge(kernel='rbf', alpha=0.05, gamma=0.1)



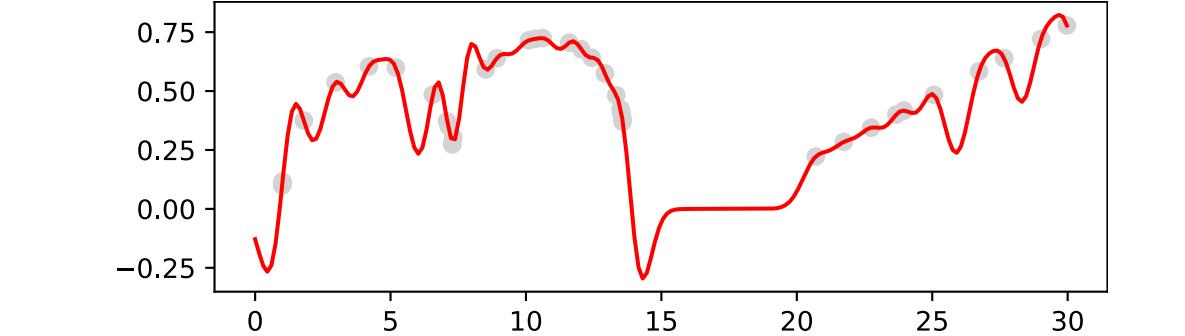
根拠なし

KernelRidge(kernel='rbf', alpha=1e-4, gamma=0.1)



根拠なし

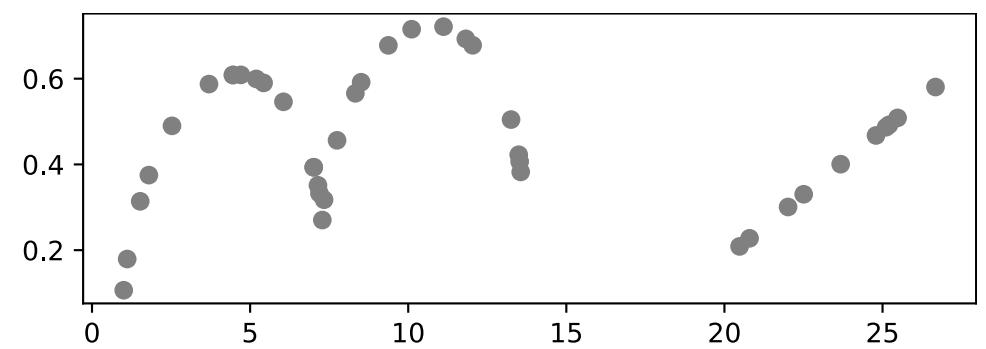
KernelRidge(kernel='rbf', alpha=1e-4, gamma=2.0)



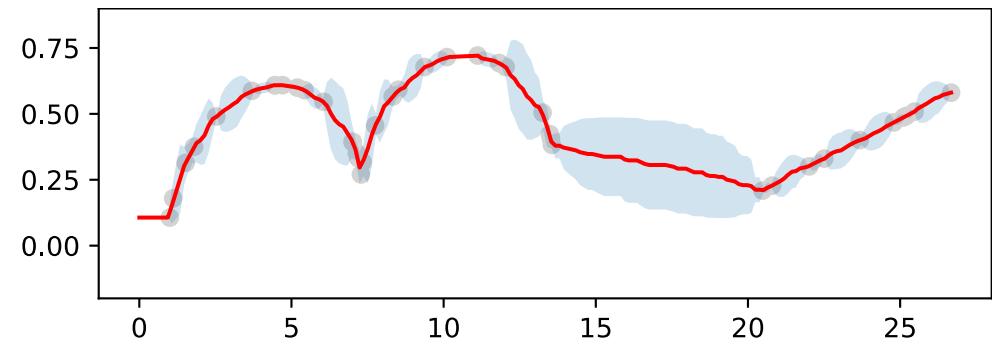
根拠なし

- この意図しない挙動はモデルの帰納バイアスに由来し検出が難しい（データがないので！）
- 材料科学ではちょっとした違いで性能が大きく改善/劣化しうるため、応答局面の連続性も特に仮定できない（Activity Cliffs, Selectivity Cliffs, ...）

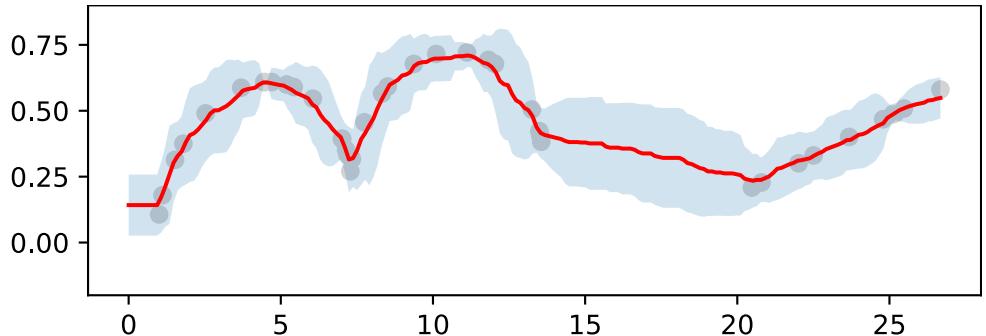
データ領域外の挙動 (UQ)



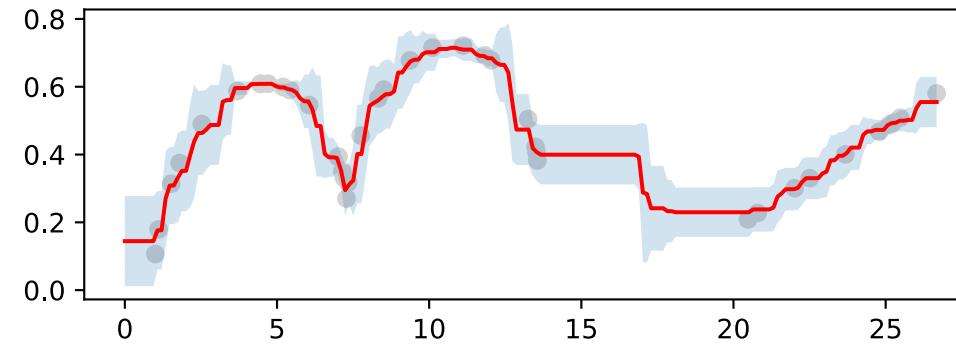
skopt.learning.
ExtraTreesRegressor(n_estimators=50)



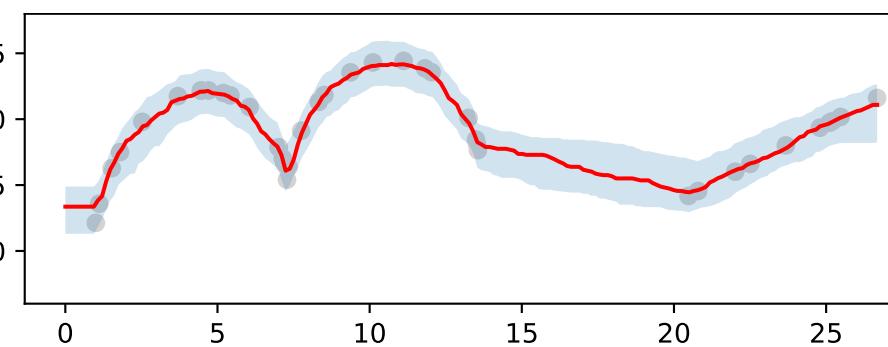
skopt.learning.
ExtraTreesRegressor(n_estimators=50,
bootstrap=True)



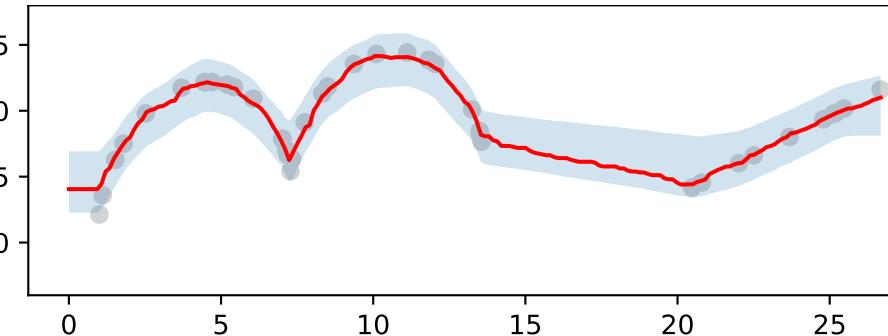
skopt.learning.
RandomForestRegressor(n_estimators=50)



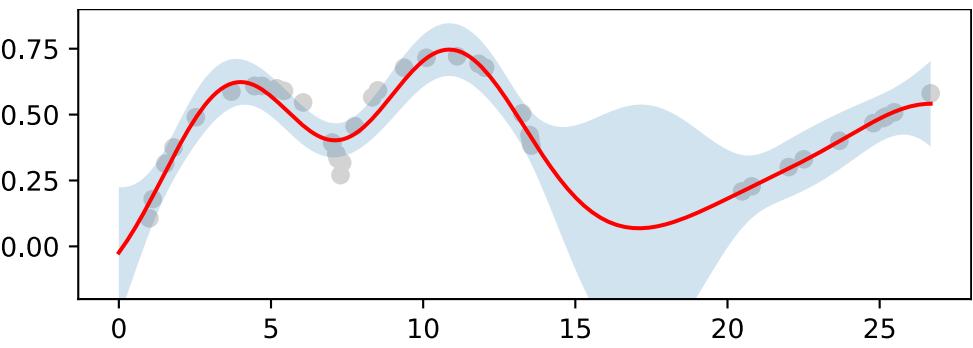
base = ExtraTreesRegressor(n_estimators=50,
bootstrap=True)
MapieRegressor(base, method="plus", cv=-1)



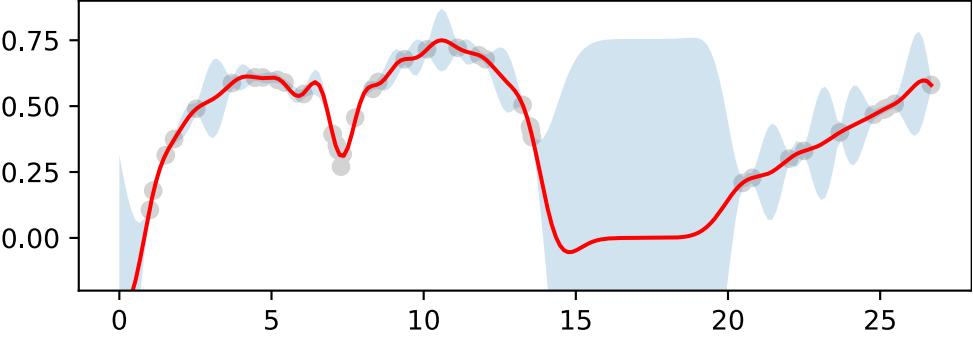
MapieRegressor(base, method="plus",
cv=Subsample(n_resamplings=50))



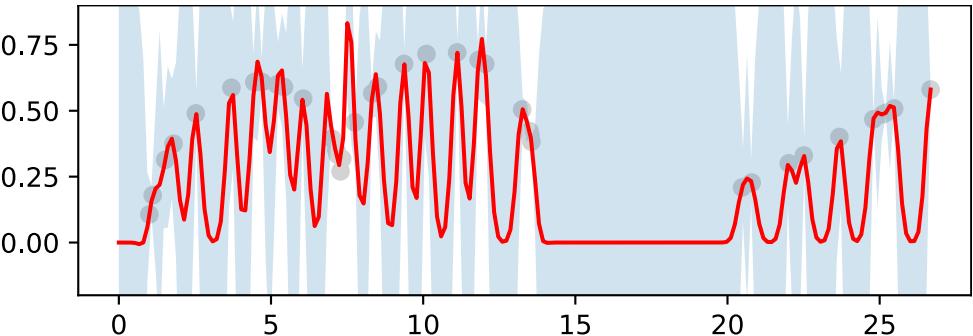
GaussianProcessRegressor(kernel=1*RBF(),
alpha=1e-2)



GaussianProcessRegressor(kernel=1*RBF(),
alpha=1e-4)



GaussianProcessRegressor(kernel=1*RBF(),
alpha=1e-5)

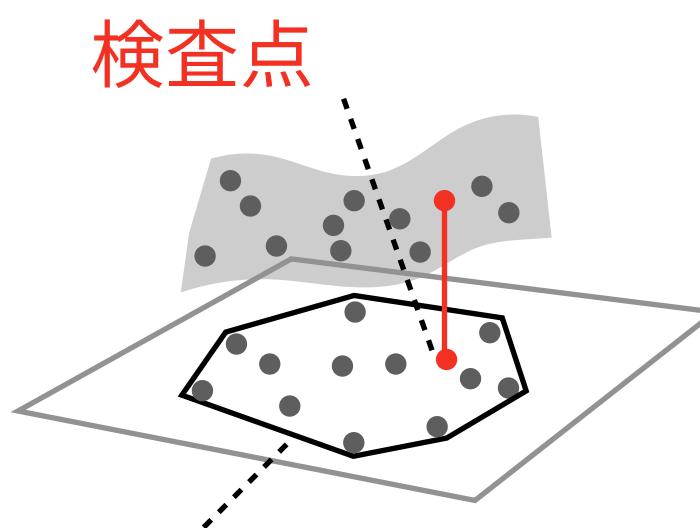


悲報：高次元では任意の検査点はほぼデータ領域外になる

高次元(>100)では訓練点のなす凸包の中に点が落ちる確率はほぼゼロで、任意の点はほぼ確実に 「データ領域外(凸包の外)」 に落ちる。(高次元機械学習はほぼ外挿!?)

悲報：高次元では任意の検査点はほぼデータ領域外になる

高次元(>100)では訓練点のなす凸包の中に点が落ちる確率はほぼゼロで、任意の点はほぼ確実に 「データ領域外(凸包の外)」 に落ちる。(高次元機械学習はほぼ外挿!?)



見本点集合
の凸包

内挿(interpolation)を「検査点が訓練データ点の凸包に落ちたときに周りの点の値からそのy値を決めること」と定義すると…

- “Our goal in this paper is to demonstrate both theoretically and empirically for both synthetic and real data that *interpolation almost surely never occurs in high-dimensional spaces (>100) regardless of the underlying intrinsic dimension of the data manifold.*”
- “Those results challenge the validity of our current interpolation/extrapolation definition as an indicator of generalization performances.”

Balestriero R, Pesenti J, LeCun Y.
[Learning in High Dimension Always Amounts to Extrapolation.](#)
arXiv [cs.LG]. 2021. <http://arxiv.org/abs/2110.09485>

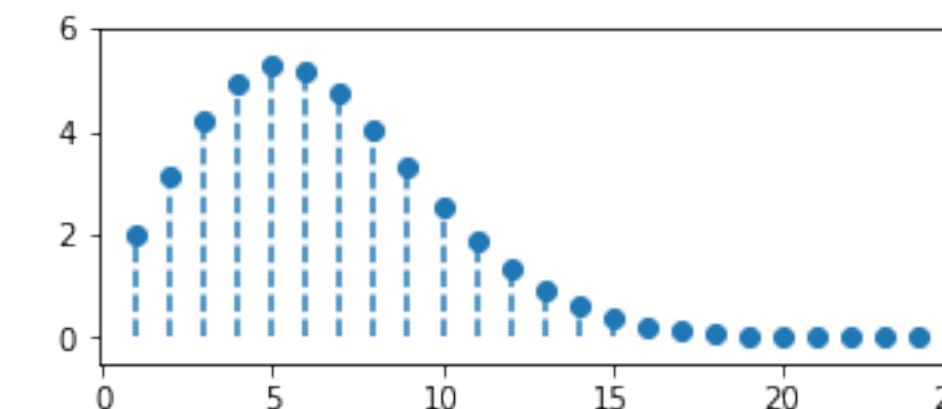
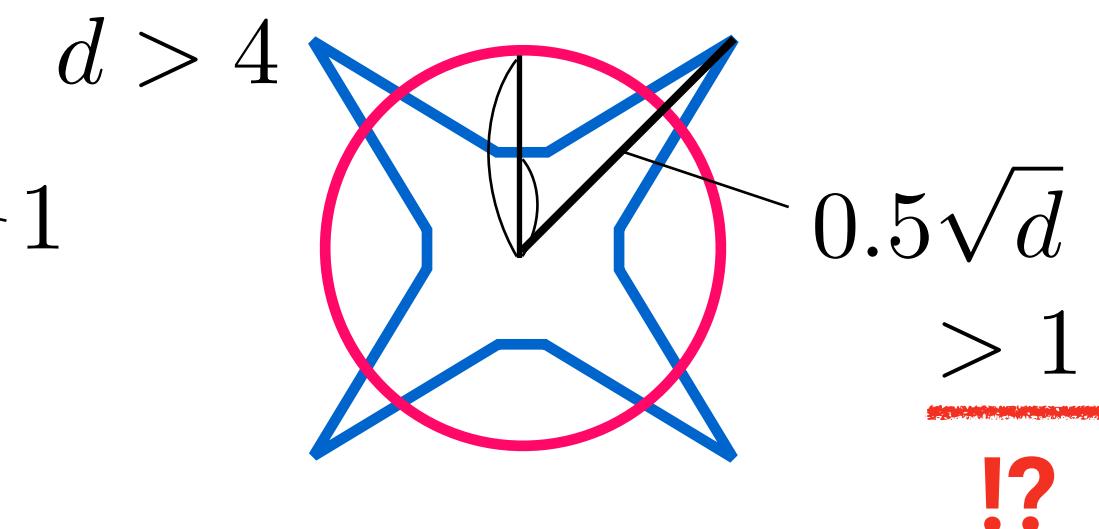
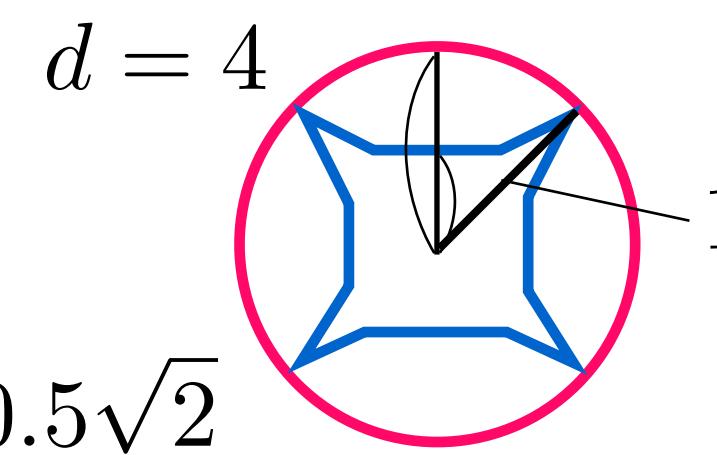
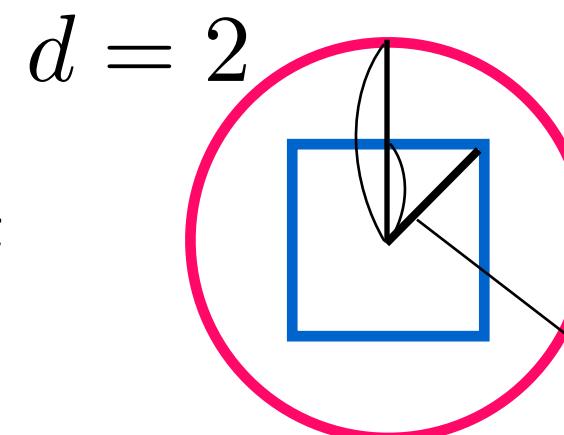
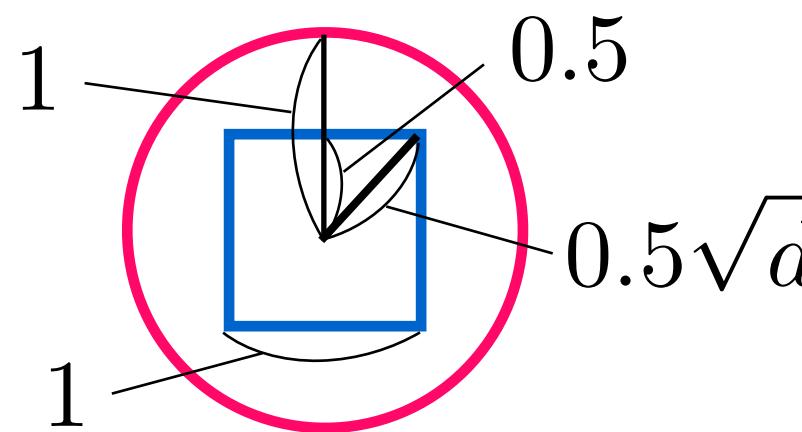
次元の呪い再び：高次元内挿/探索の困難性・非直観性 1/2

球(ある点から等距離にある点集合)の中身は次元を大きくすると急激に0に縮退していくため
高次元空間ではある点の近くに別の点を取るのは一様な確率探索ではほぼ有り得ない

d次元空間における単位球($r=1$ の超球)の体積

$$V(d) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2} \cdot \Gamma\left(\frac{d}{2}\right)} \rightarrow 0 \quad (d \rightarrow \infty)$$

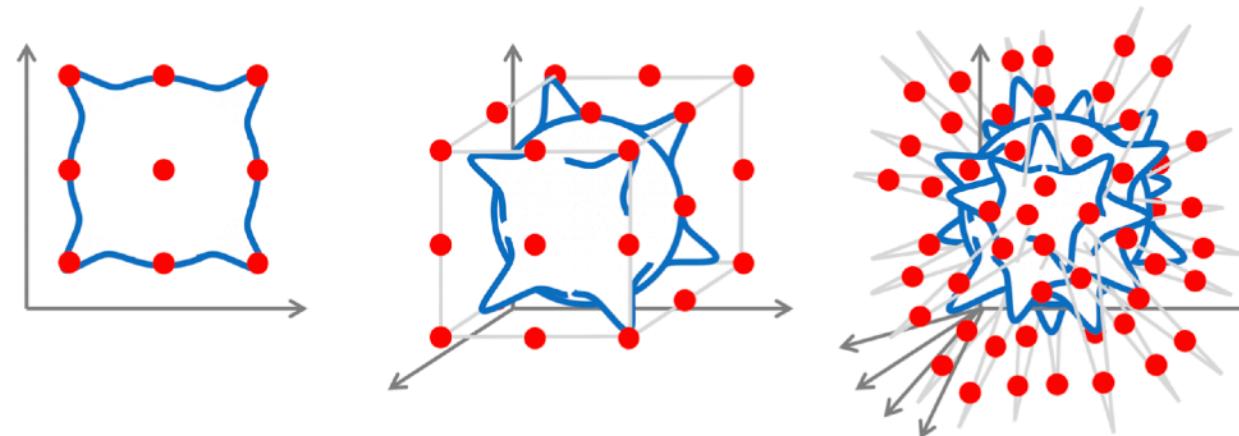
Unit Cube in
Unit Ball



次元の呪い再び：高次元内挿/探索の困難性・非直観性 2/2

もし似た入力には似た出力をという緩い連續性(Lipschitz連續性)しか課さない場合には一様な ε 近似に $(1/\varepsilon)^d$ オーダのサンプル数が必要になる。Sobolev classとかにしても良くならない。

例： $d=2$ で「100点くらい」の精度を求めるなら $d=10$ では $10^{10}=100$ 億点必要で非現実的…



$$L\text{-Lipschitz } f : \mathbb{R}^d \rightarrow \mathbb{R}$$
$$|f(x) - f(x')| \leq L\|x - x'\| \quad \text{for all } x, x' \in \mathbb{R}^d$$

Donoho DL,

High-dimensional data analysis: The curses and blessings of dimensionality.

Plenary Lecture, AMS National Meeting on Mathematical Challenges of the 21st Century. 2000.

Bronstein MM, Bruna J, Cohen T, Veličković P.

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

arXiv [cs.LG]. 2021. <http://arxiv.org/abs/2104.13478>

現実的帰結：UnderspecificationとRashomon効果

Underspecification

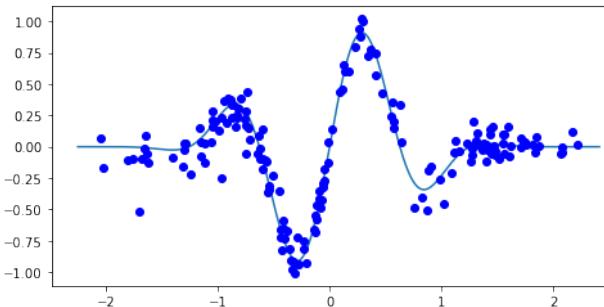
高次元の機械学習では、全域での関数近似を保証するには(指数的広さのモデル探索空間をspecifyするには)本質的にほとんど常にデータが足りていない

羅生門効果：良い機械学習モデルの多重性 (非一意性)

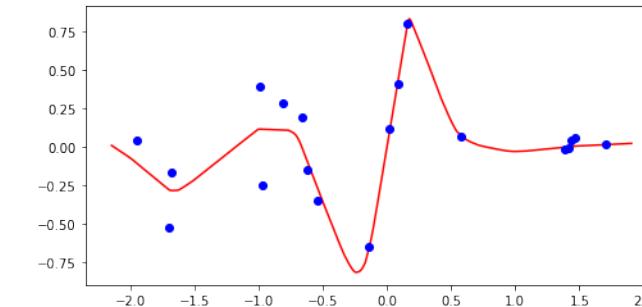
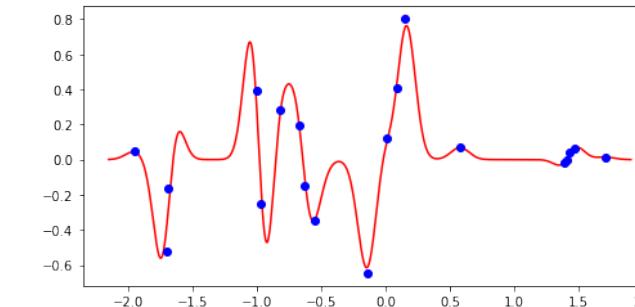
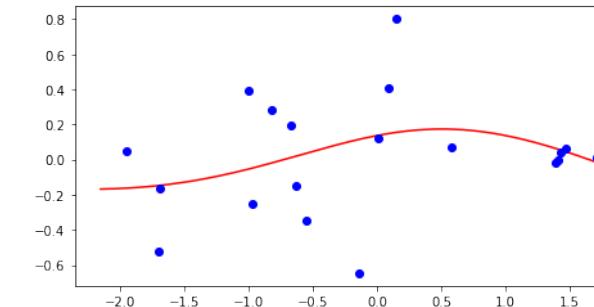
同程度にCV精度が良い機械学習モデルは一つのデータセットからたくさん作れる

少数データではより顕著だが、「ビッグデータ」でも高次元ではスカスカでこの現象が起こる。

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



D'Amour et al., Underspecification Presents Challenges for Credibility in Modern Machine Learning.

J Mach Learn Res, 2022; 23: 1-61. <https://ai.googleblog.com/2021/10/how-underspecification-presents.html>

現実的帰結：UnderspecificationとRashomon効果

Underspecification

高次元の機械学習では、全域での関数近似を保証するには(指数的広さのモデル探索空間をspecifyするには)本質的にほとんど常にデータが足りていない



というか、考えうる無数の候補から有望なものをspecifyできるデータが足りておらず確証が持てないので「機械学習というやつでも使ってみよう」となっているのだと思われる。

現実的帰結：UnderspecificationとRashomon効果

Underspecification

高次元の機械学習では、全域での関数近似を保証するには(指数的広さのモデル探索空間をspecifyするには)本質的にほとんど常にデータが足りていない



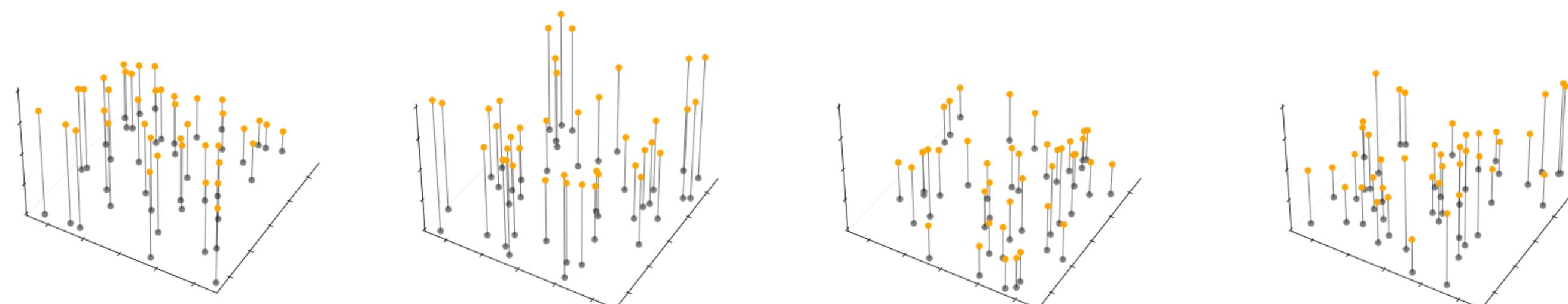
というか、考えうる無数の候補から有望なものをspecifyできるデータが足りておらず確証が持てないので「機械学習というやつでも使ってみよう」となっているのだと思われる。

自然科学でのデータ利活用においてこの**情報の部分性(Underspecification)**は本質的的前提もし上の意味で、十分なデータがある状況であれば、機械学習なんか使わなくとも既に良い発見が得られているはずであるし、機械学習を使って得られるプラスαはほぼ皆無？

データ領域外の挙動が支配的な影響を与える！

探索が目的だとすると「高次元」の任意の点での予測が求められるが、有限の見本点は(“big data”としても)高次元空間ではスカスカで「データ領域外」での挙動が支配的な影響を与える。(ほぼすべての検査点がデータ領域外になる)

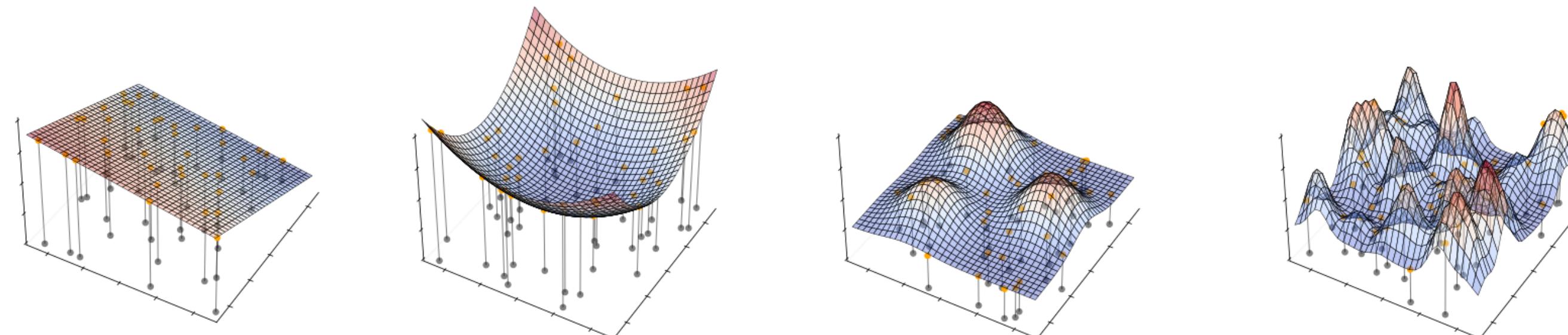
「機械学習を活用してみよう」というとき、普通は近似しようとする関数関係が複雑であり、2次元でも無理ゲーぽいのに高次元では見本例に基づく近似はほぼ絶望的では…
→ 「データ領域外」で根拠のない妙な挙動が起こらない超保守的で安全な予測手法が良さそう



データ領域外の挙動が支配的な影響を与える！

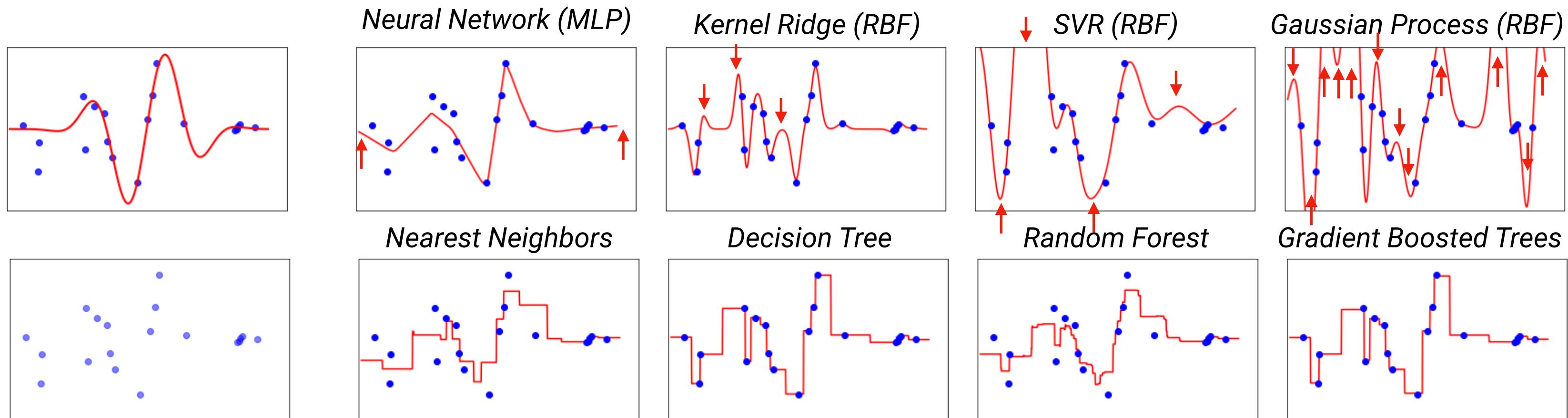
探索が目的だとすると「高次元」の任意の点での予測が求められるが、有限の見本点は(“big data”としても)高次元空間ではスカスカで「データ領域外」での挙動が支配的な影響を与える。(ほぼすべての検査点がデータ領域外になる)

「機械学習を活用してみよう」というとき、普通は近似しようとする関数関係が複雑であり、2次元でも無理ゲーぽいのに高次元では見本例に基づく近似はほぼ絶望的では…
→ 「データ領域外」で根拠のない妙な挙動が起こらない超保守的で安全な予測手法が良さそう



超保守的で安全な予測手法？

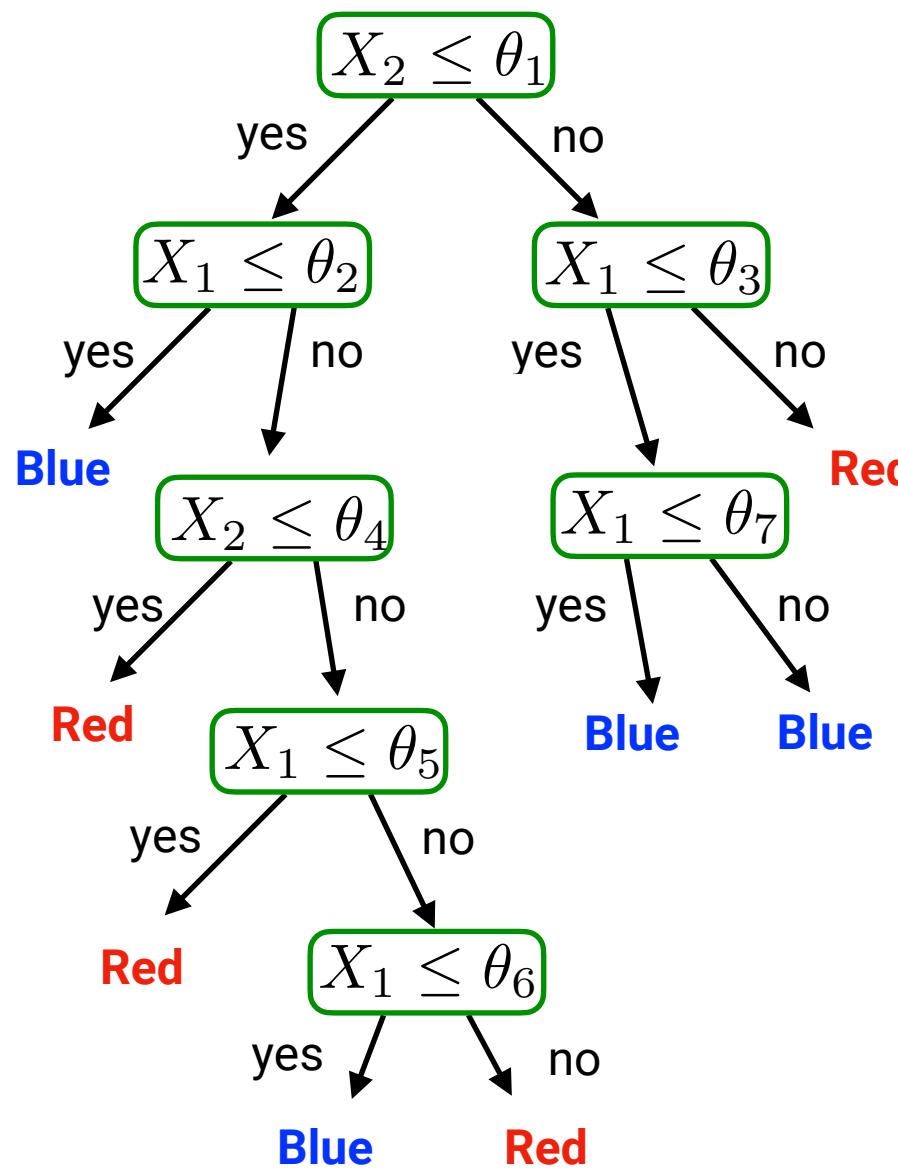
- 下記は低次元だが意図的に過適合させた例：高次元では「至るところデータ領域外」だと示唆されるので意図しない予測が起こるのはできるだけ避けたい…
- 決定木回帰やNearest Neighborは見本例と無関係な補間を原理上しないため、高次元探索のベースラインに良さそう (Neural Networkはデータ領域外で線形予測になり端の個数が指数的に増大する高次元探索では良くないかも？)



決定木=データに依存した空間分割上の区分的定数予測

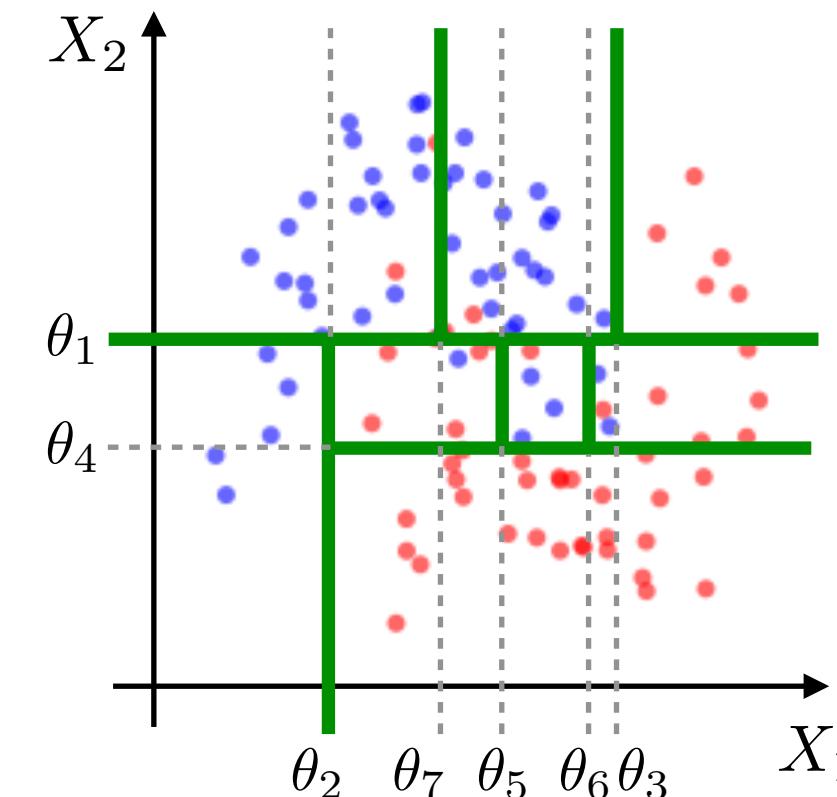
決定木は3つの側面から把握することができる

① 入れ子型のif-thenルール

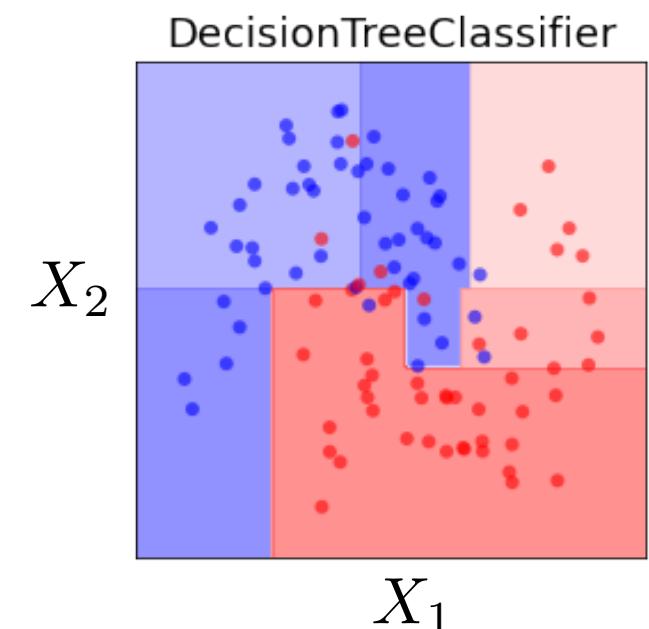


```
if x₂ ≤ θ₁ then
    if x₁ ≤ θ₂ then
        return Blue
    else
        if x₂ ≤ θ₄ then
            return Red
        else
            if x₁ ≤ θ₅ then
                return Red
            else
                if x₁ ≤ θ₆ then
                    return Blue
                else
                    return Red
            else
                if x₁ ≤ θ₃ then
                    if x₂ ≤ θ₇ then
                        return Blue
                    else
                        return Blue
                else
                    return Red
    else
        if x₁ ≤ θ₇ then
            return Blue
        else
            return Red
```

② 入力空間の再帰的分割



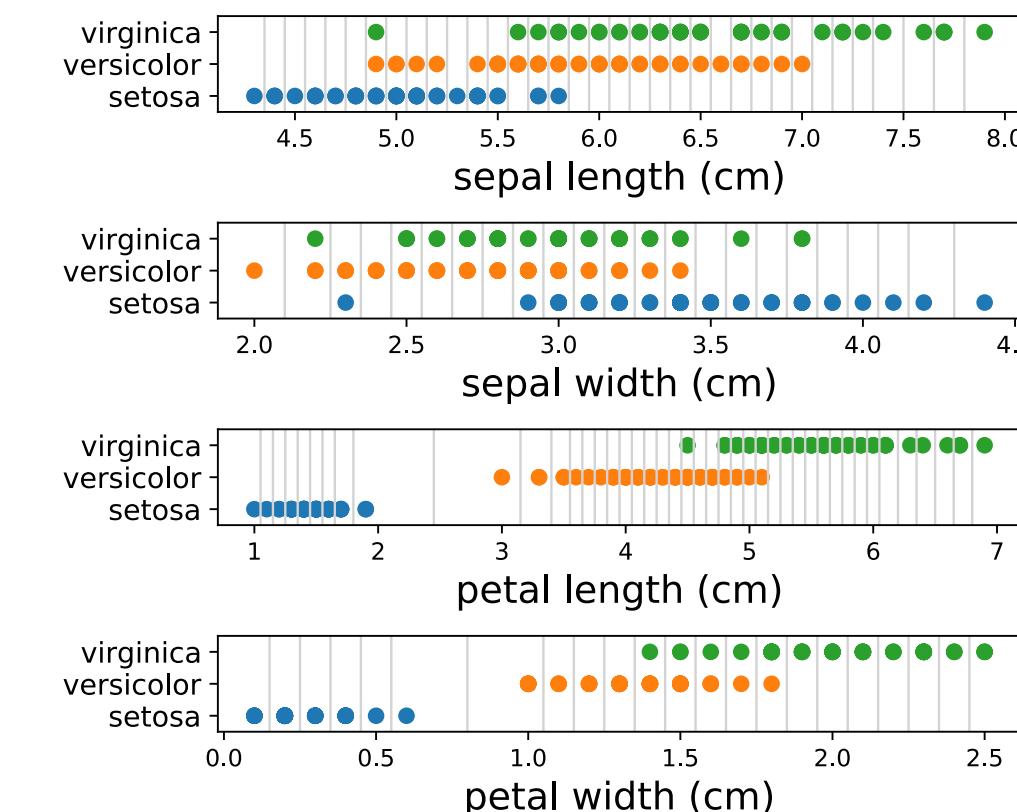
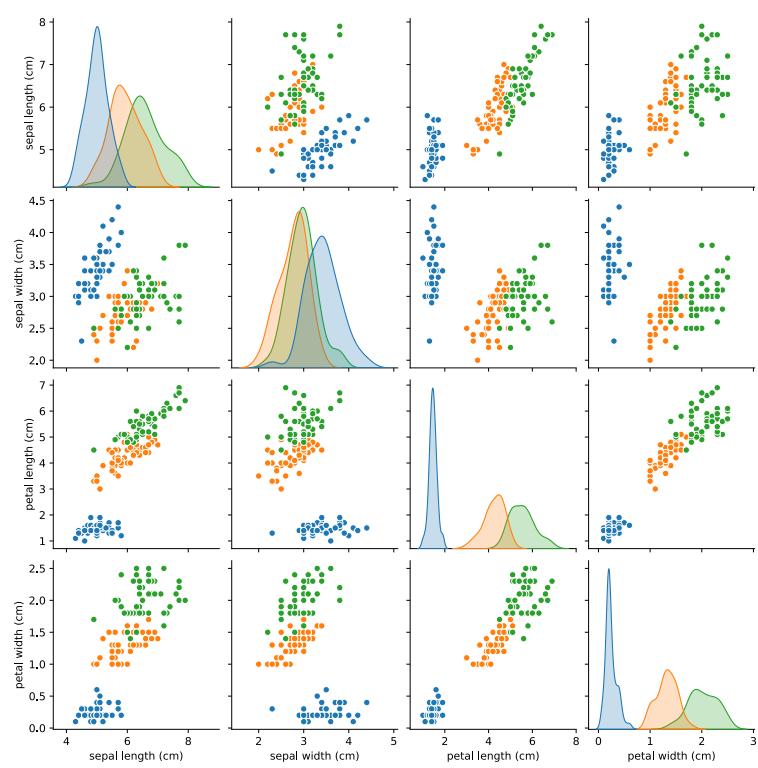
③ 区分的定数予測



決定木では多変量を同時に見るのを諦める！

「次元の呪い」は基本的に多変量を「同時に見ようとするから」起きる問題

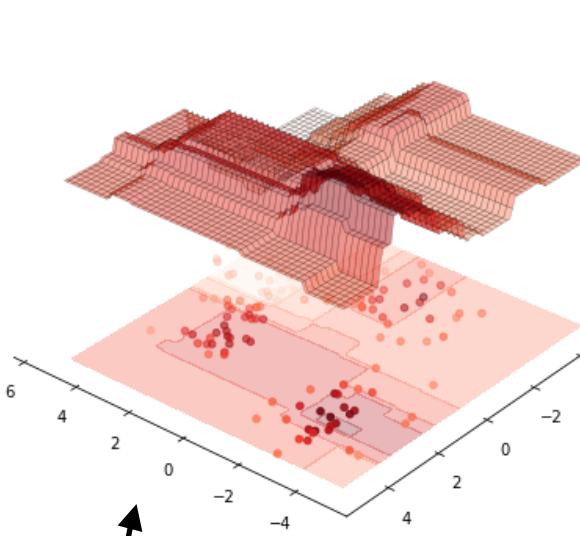
- 変数の組合せを考えると次元 d が指数で効いてしまう
- 一方、標準的な決定木は「**各軸ごとの1次元射影**」でしか多変量データを見ない
→ 変数の組合せ(交互作用)はこの分割を「再帰」することで考慮される！



決定木アンサンブル(予測精度・平滑性・不安定性を改善！)

複数の決定木の加法的アンサンブル

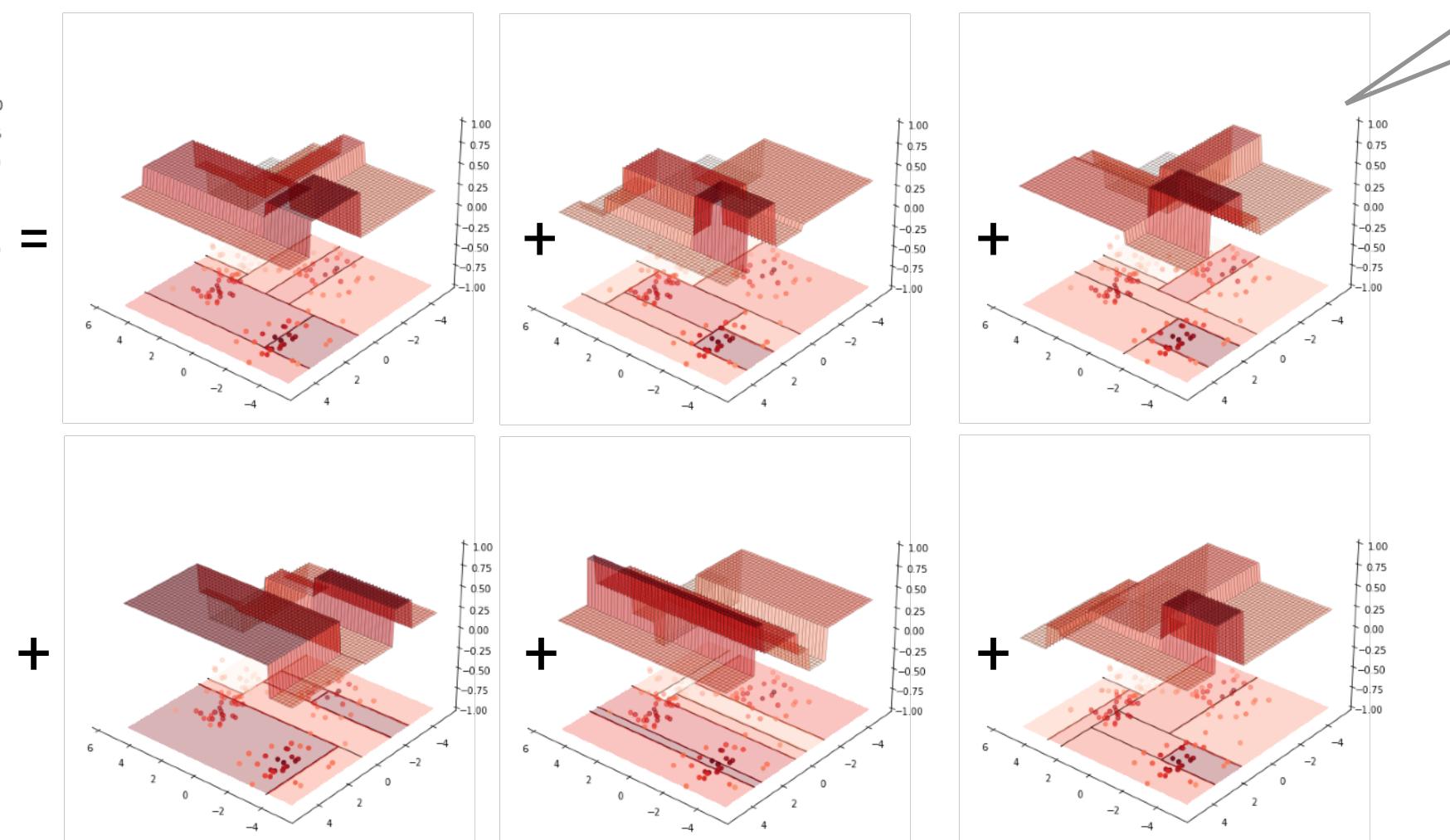
RandomForestRegressor
(n_estimators=6,
max_leaf_nodes=8)



アンサンブル後の
領域数は組合せで
8より大幅に増える！

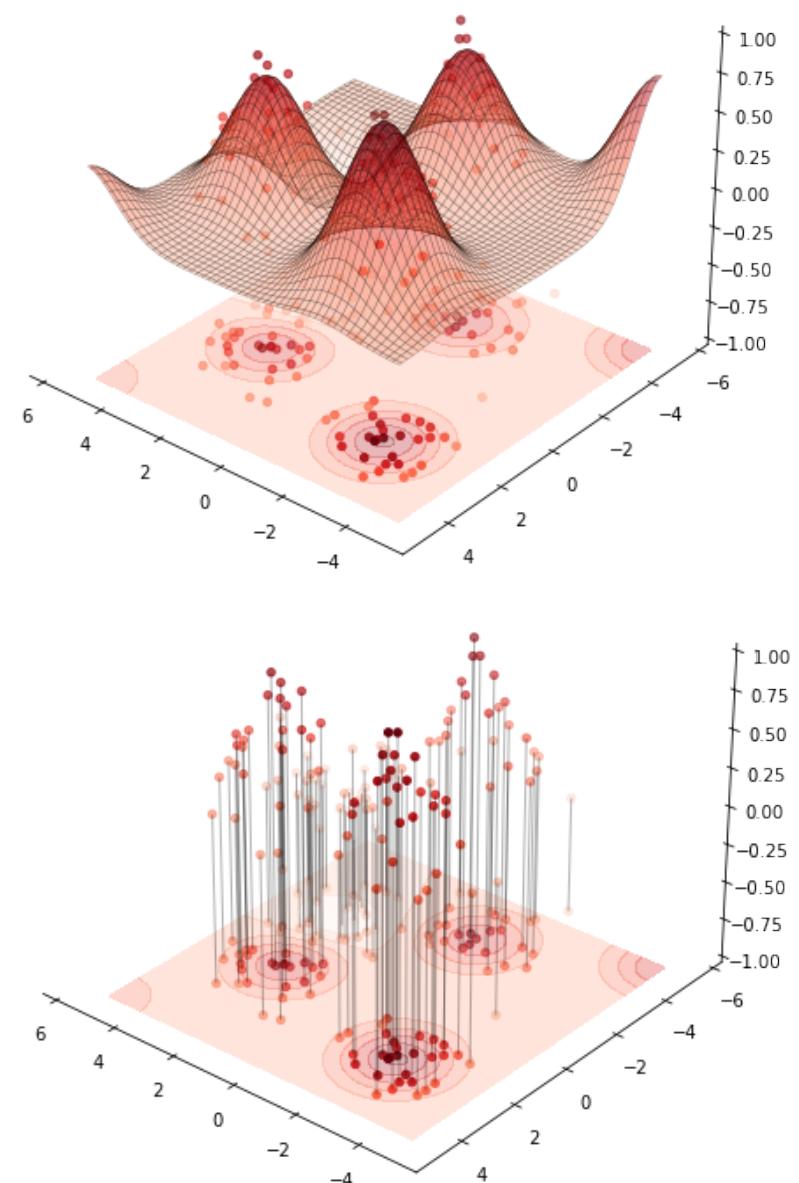
$6 \times \text{DecisionTreeRegressor}(\text{max_leaf_nodes}=8)$

必要なら `min_samples_leaf` で領域のサンプル数の下限を設定



領域数を8に設定

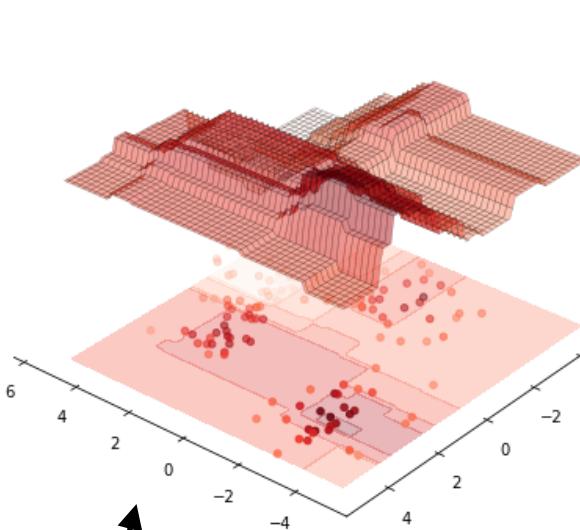
領域分割上のヒストグラム近似
(区分定数予測)



決定木アンサンブル(予測精度・平滑性・不安定性を改善！)

複数の決定木の加法的アンサンブル

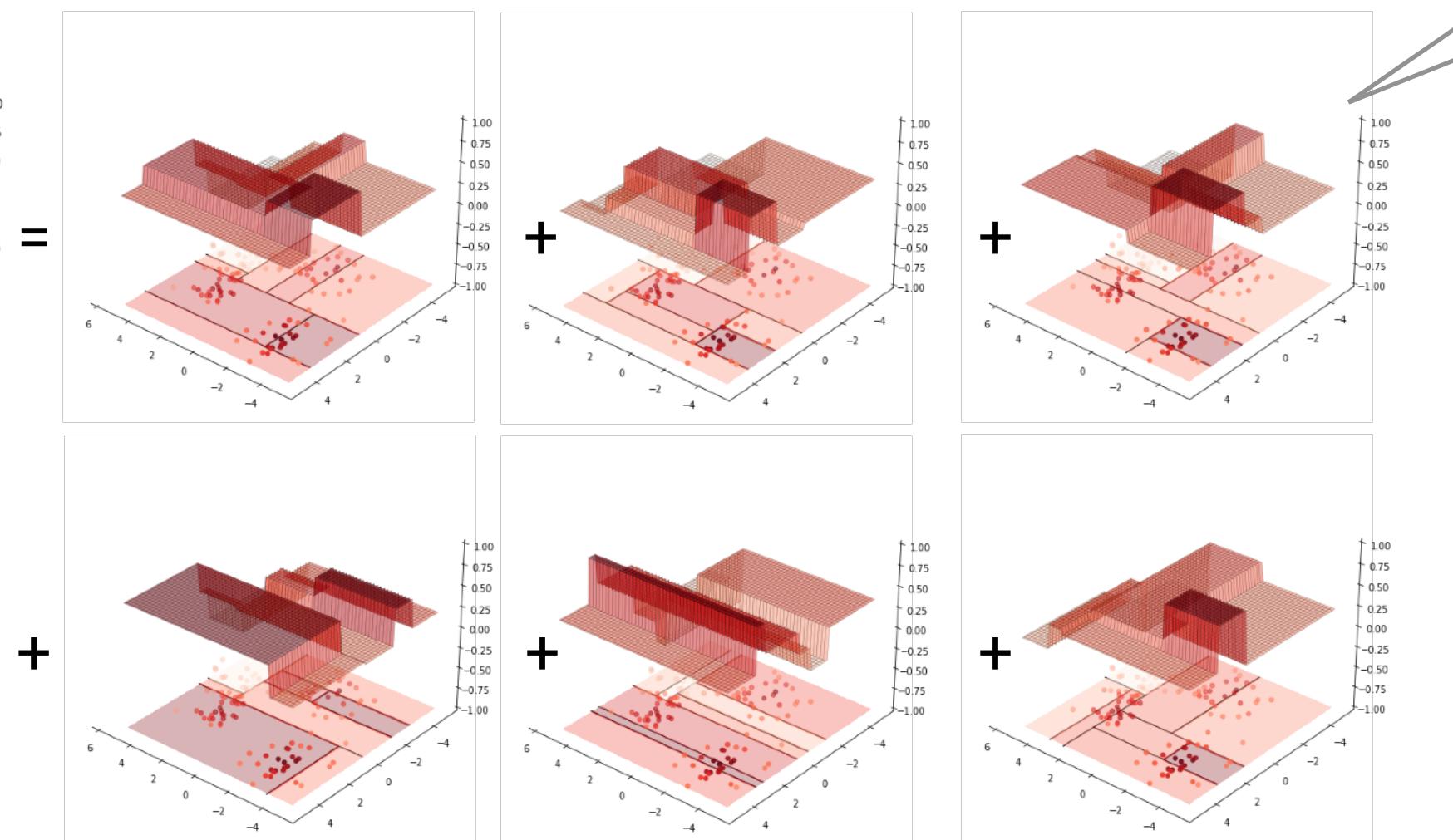
RandomForestRegressor
(n_estimators=6,
max_leaf_nodes=8)



アンサンブル後の
領域数は組合せで
8より大幅に増える！

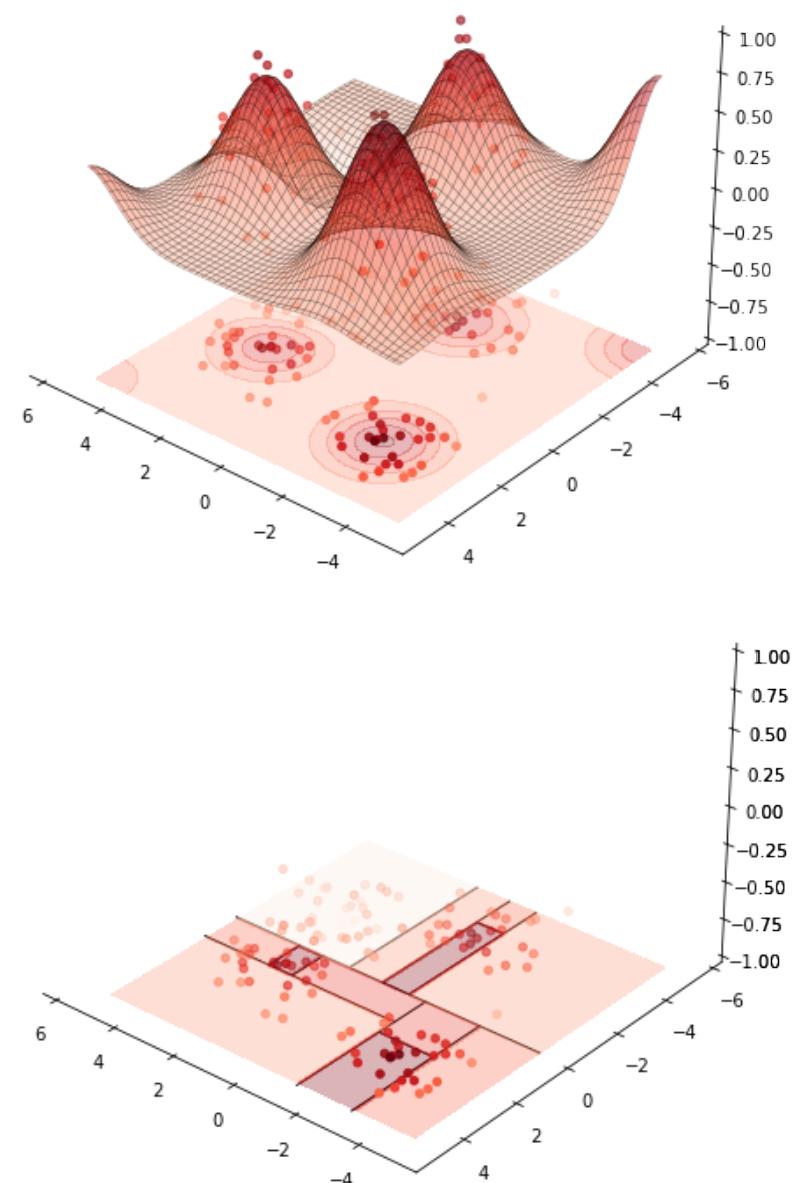
$6 \times \text{DecisionTreeRegressor}(\text{max_leaf_nodes}=8)$

必要なら `min_samples_leaf` で領域のサンプル数の下限を設定



領域数を8に設定

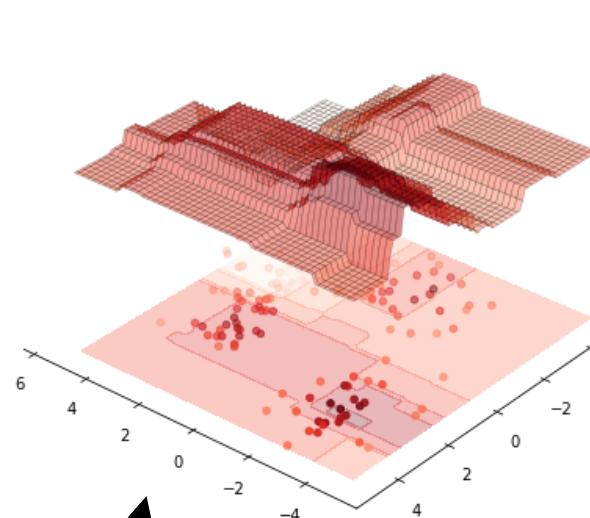
領域分割上のヒストグラム近似
(区分定数予測)



決定木アンサンブル(予測精度・平滑性・不安定性を改善！)

複数の決定木の加法的アンサンブル

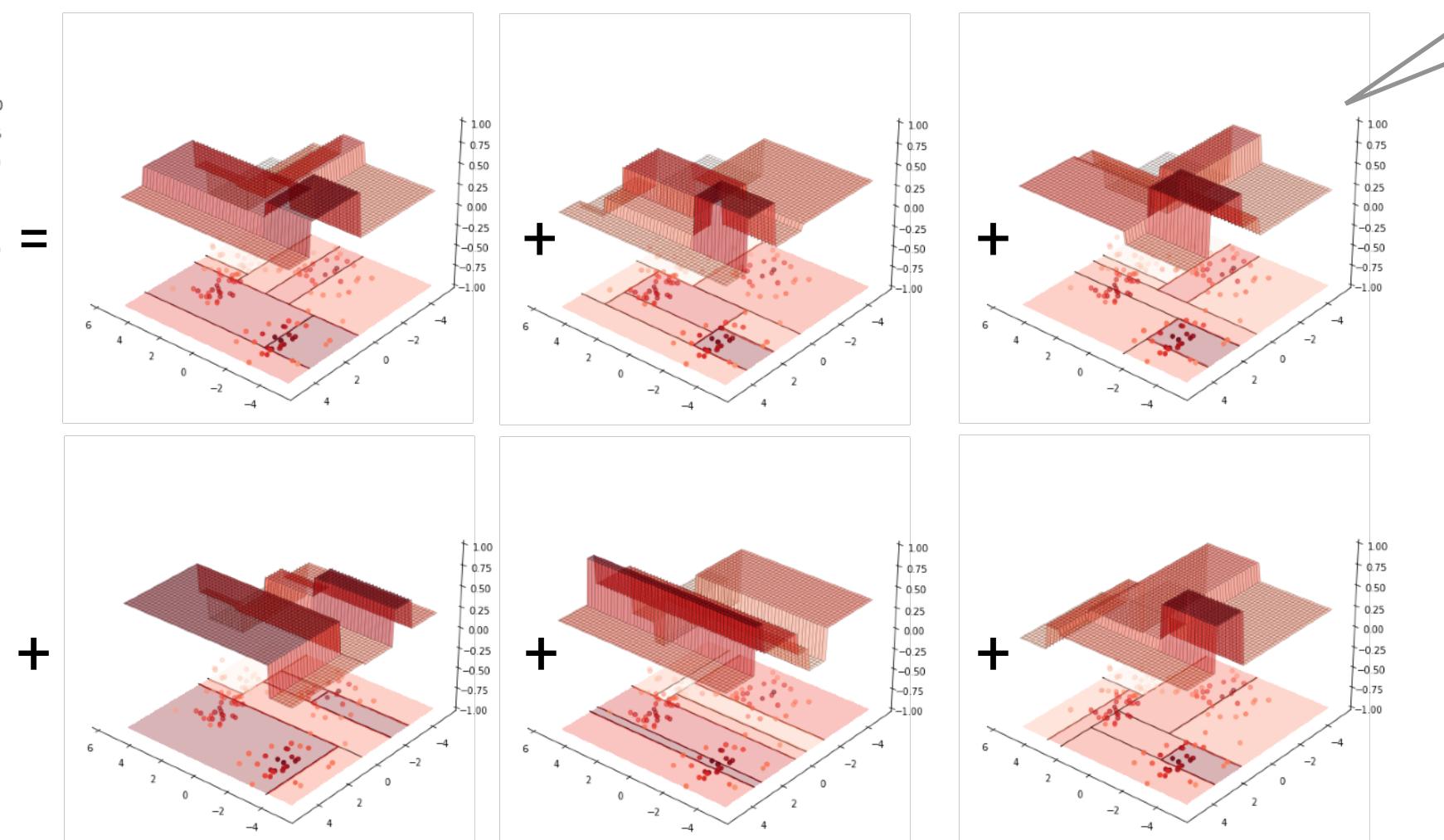
RandomForestRegressor
(n_estimators=6,
max_leaf_nodes=8)



アンサンブル後の
領域数は組合せで
8より大幅に増える！

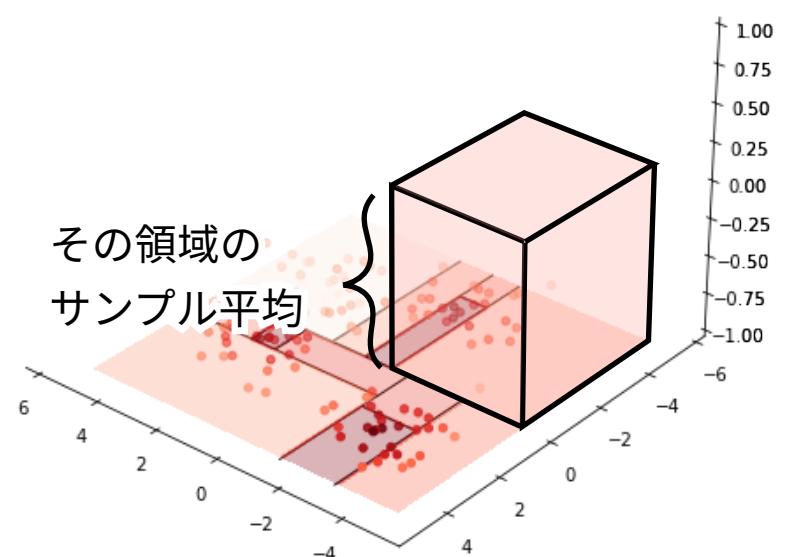
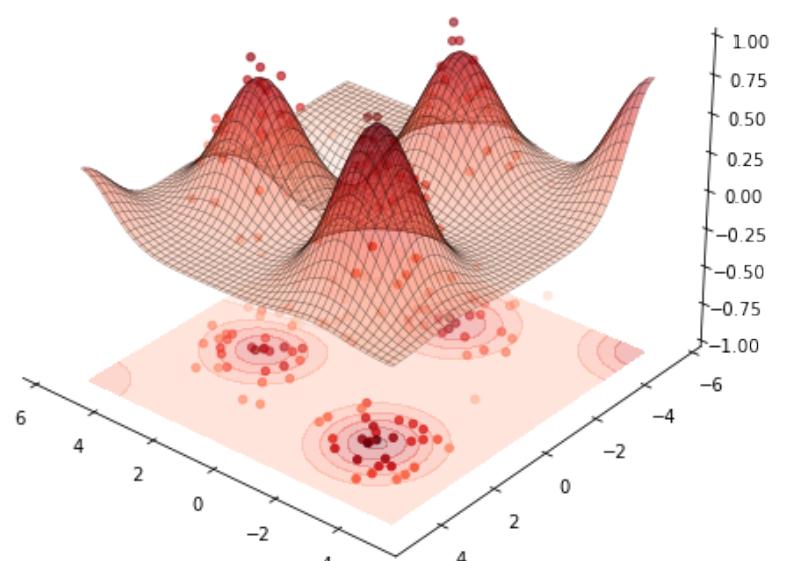
$6 \times \text{DecisionTreeRegressor}(\text{max_leaf_nodes}=8)$

必要なら `min_samples_leaf` で領域のサンプル数の下限を設定



領域数を8に設定

領域分割上のヒストグラム近似
(区分定数予測)

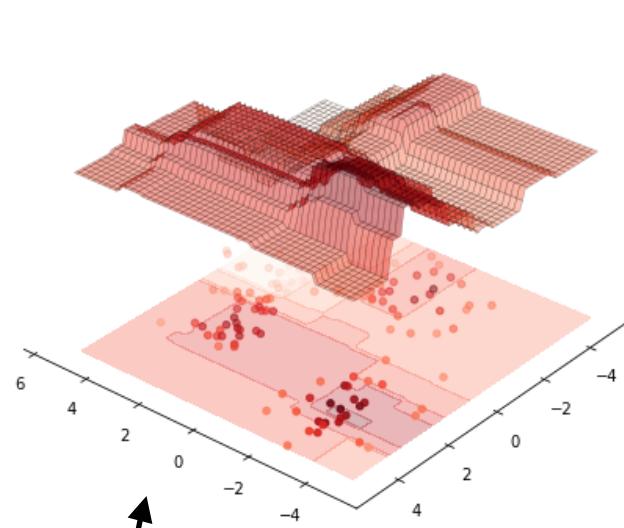


その領域の
サンプル平均

決定木アンサンブル(予測精度・平滑性・不安定性を改善！)

複数の決定木の加法的アンサンブル

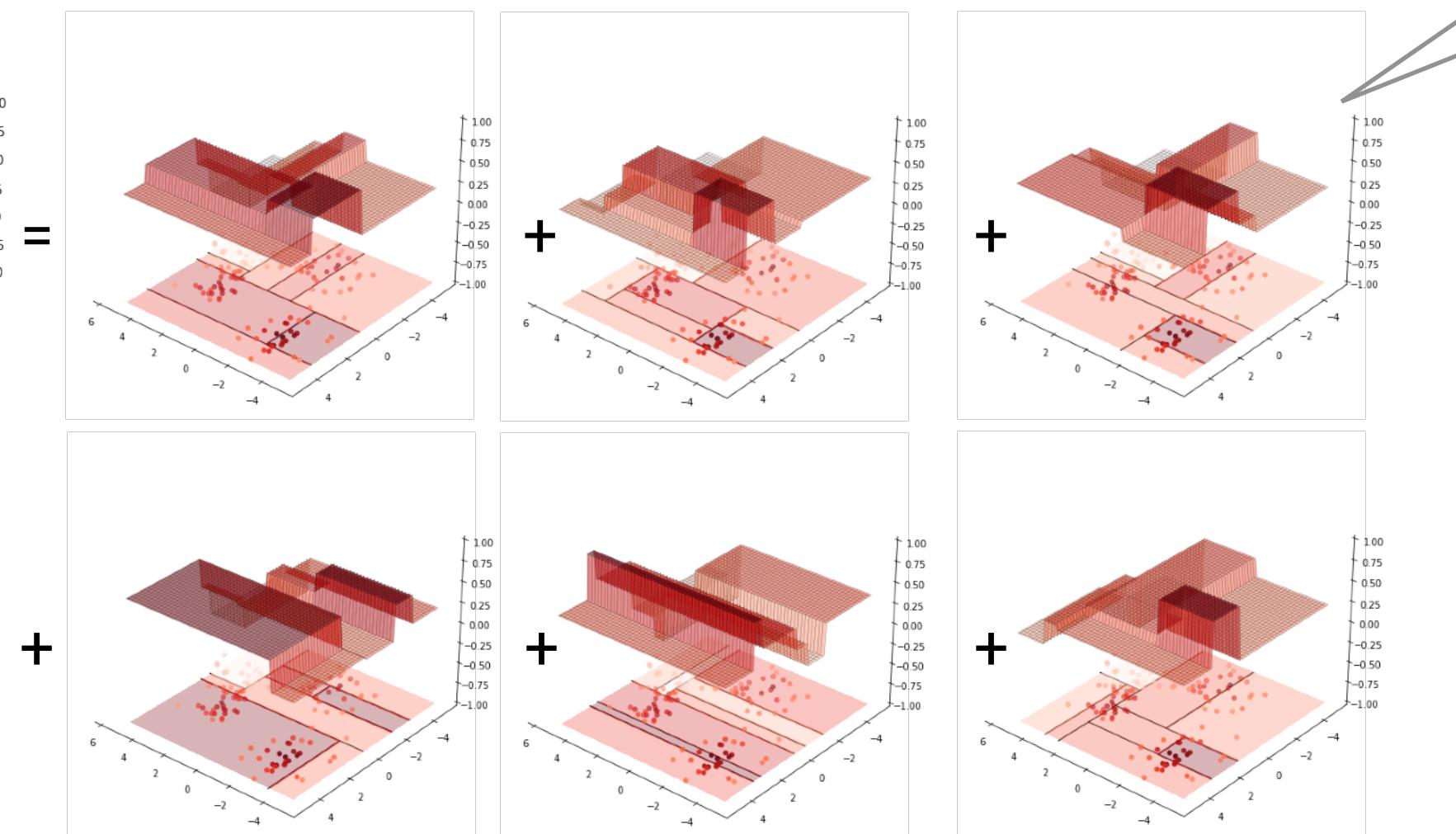
RandomForestRegressor
(n_estimators=6,
max_leaf_nodes=8)



アンサンブル後の
領域数は組合せで
8より大幅に増える！

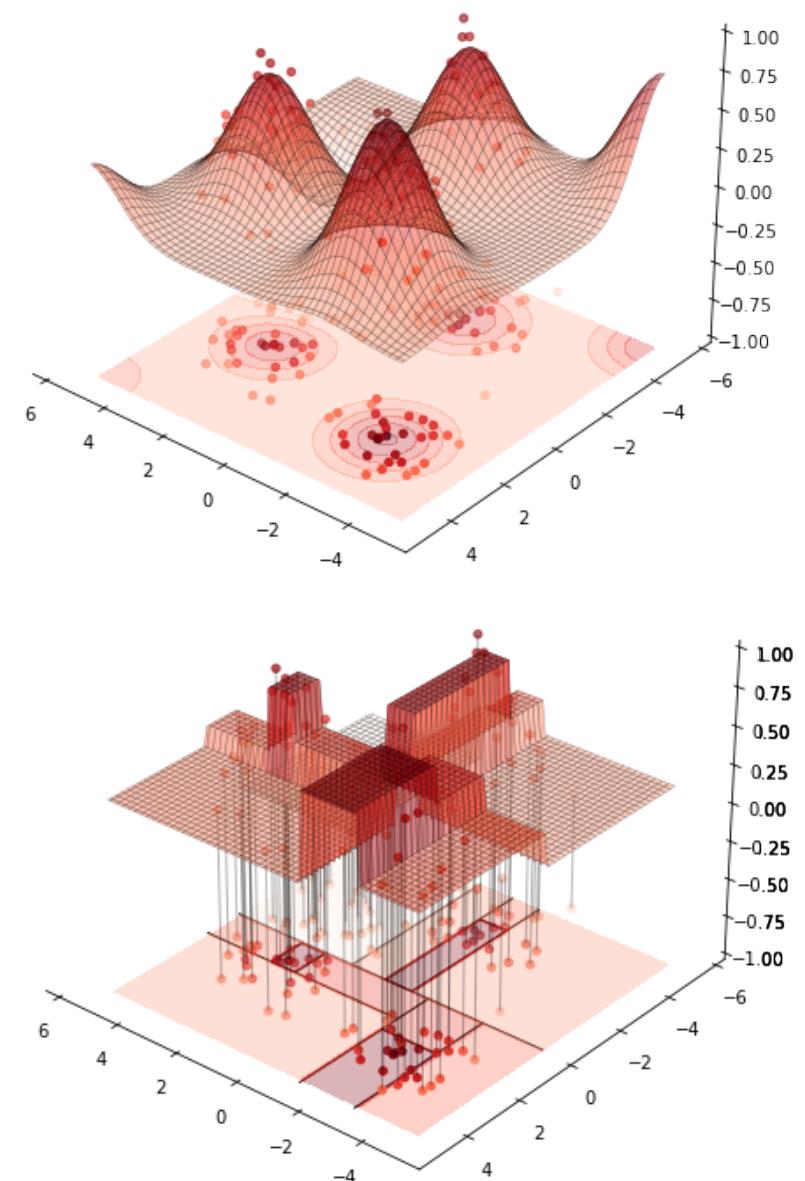
$6 \times \text{DecisionTreeRegressor}(\text{max_leaf_nodes}=8)$

必要なら `min_samples_leaf` で領域のサンプル数の下限を設定



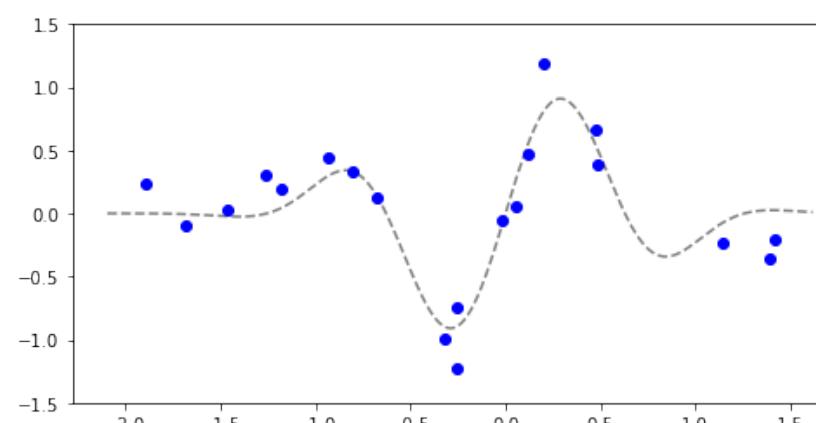
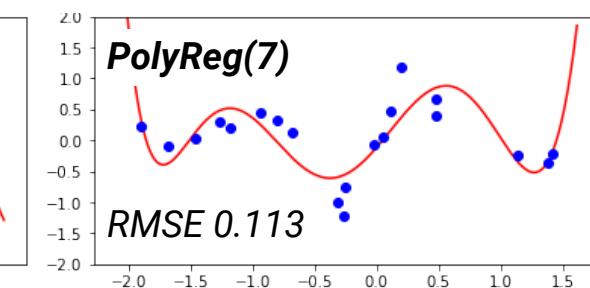
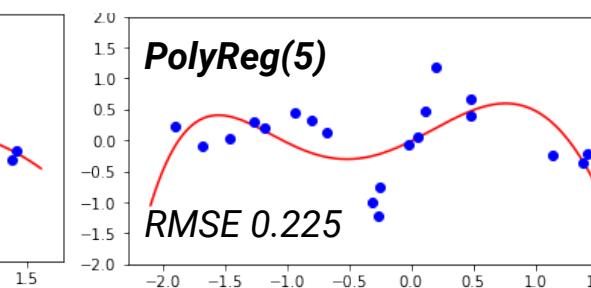
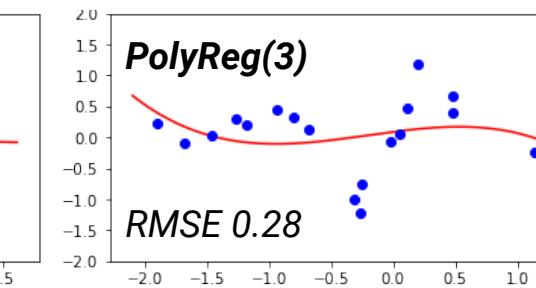
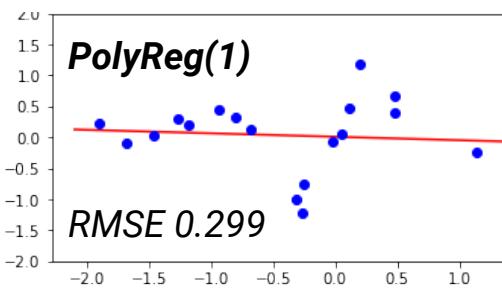
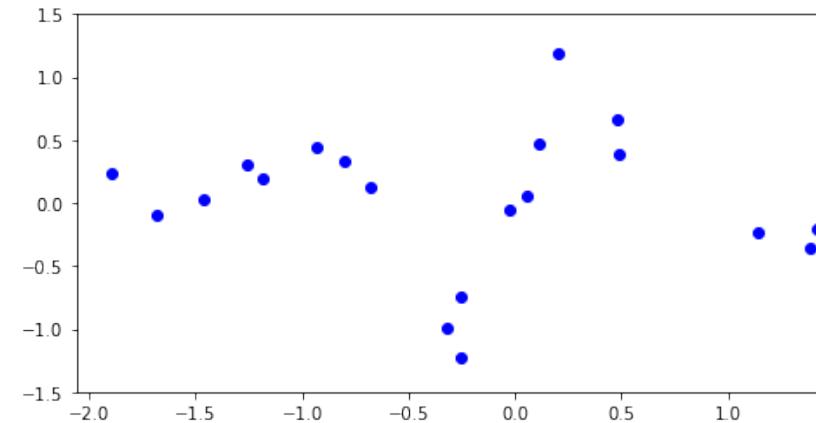
領域数を8に設定

領域分割上のヒストグラム近似
(区分定数予測)

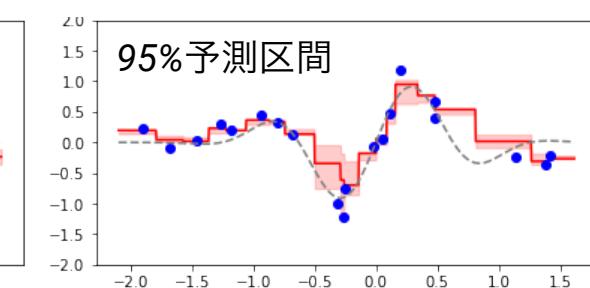
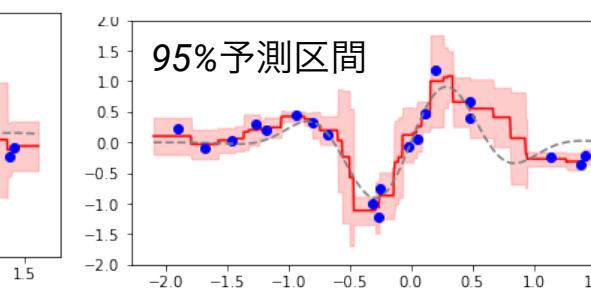
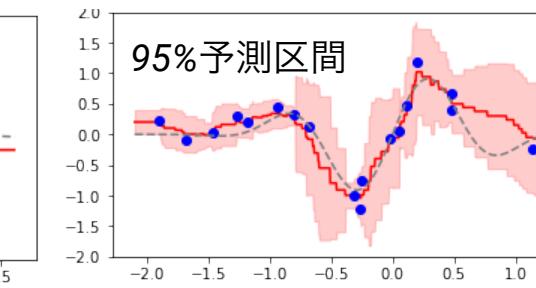
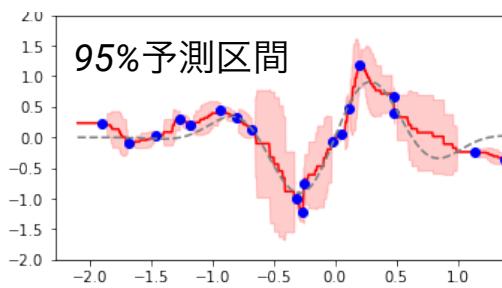
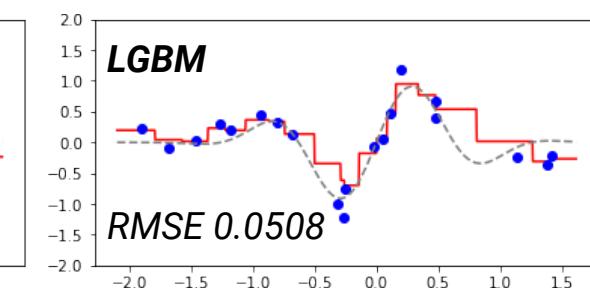
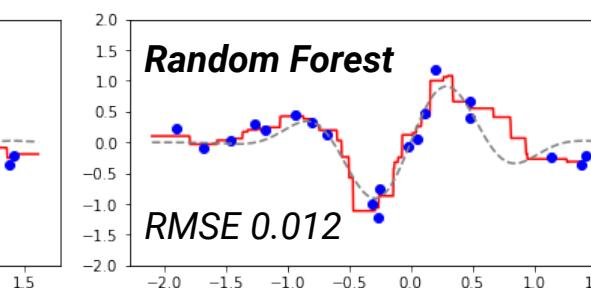
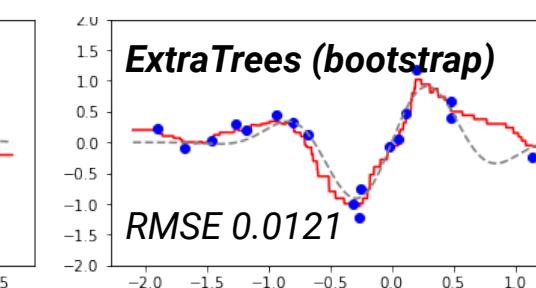
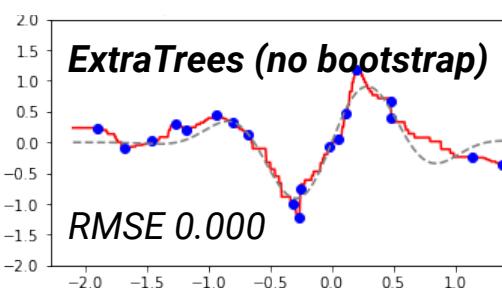


決定木アンサンブル：局所性による無害な/良性の過適合

Problematic overfitting by polynomial regression of order k

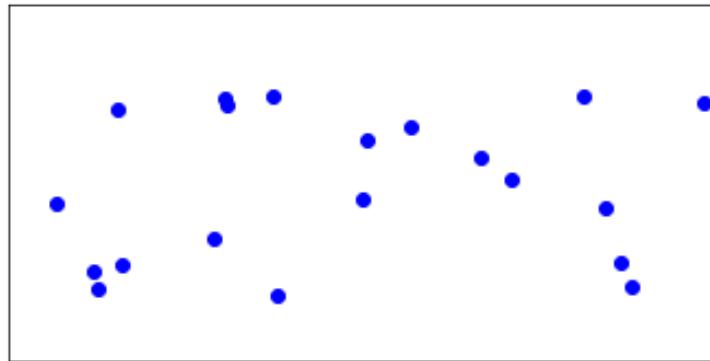


clearly overfitted but harmless (still informative)

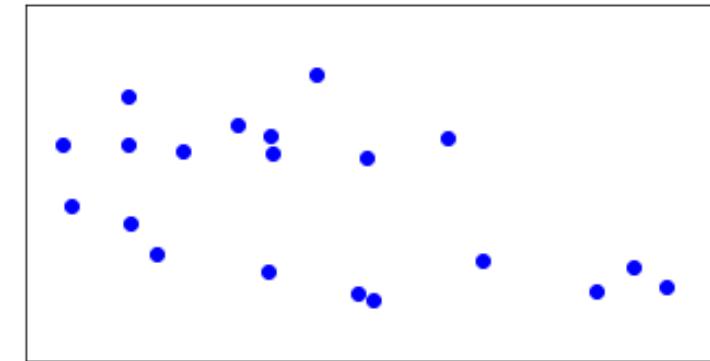


局所性をもつInterpolator

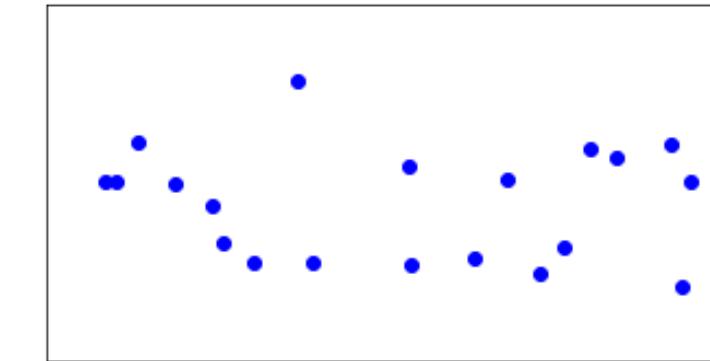
全くランダムなラベルでも訓練誤差0を達成できる (簡単そうに聞こえるが高次元では非自明)



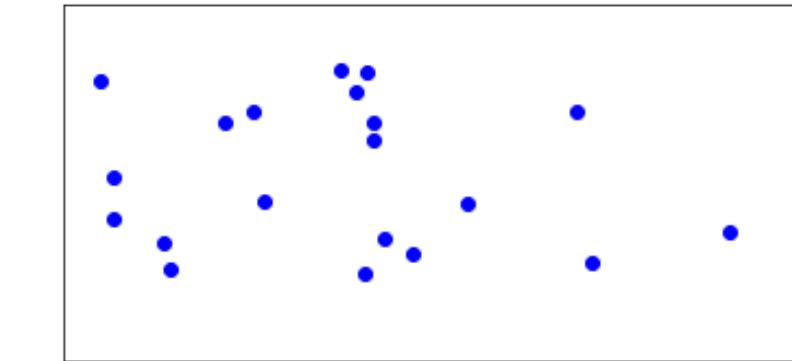
ExtraTrees (no bootstrap)



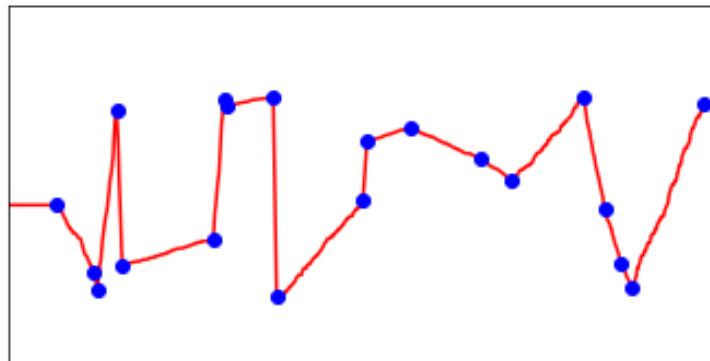
ExtraTrees (no bootstrap)



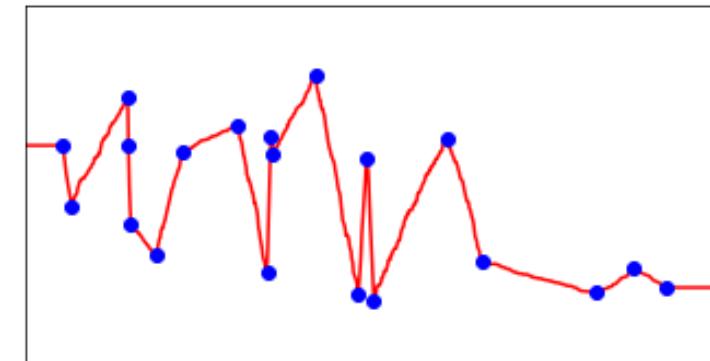
ExtraTrees (no bootstrap)



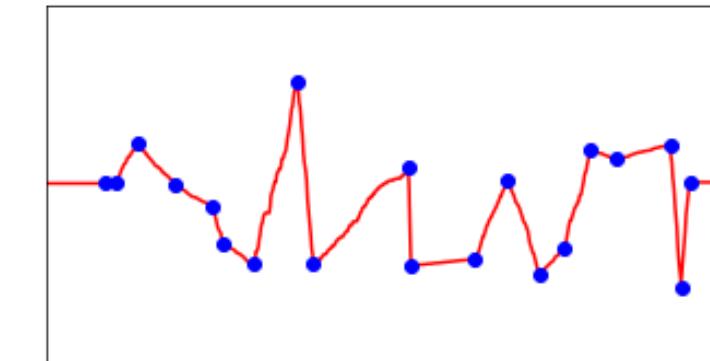
ExtraTrees (no bootstrap)



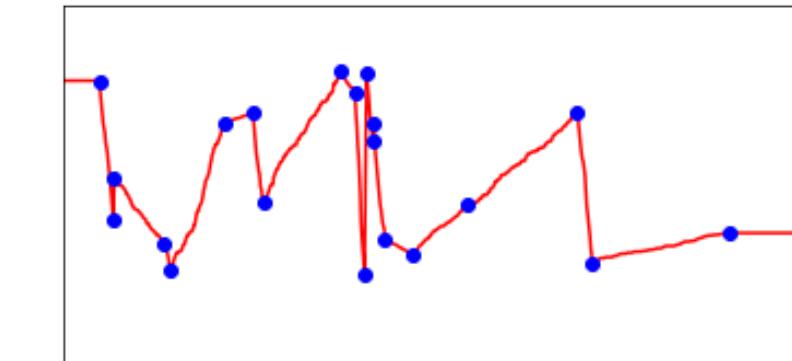
Gradient Boosted Trees



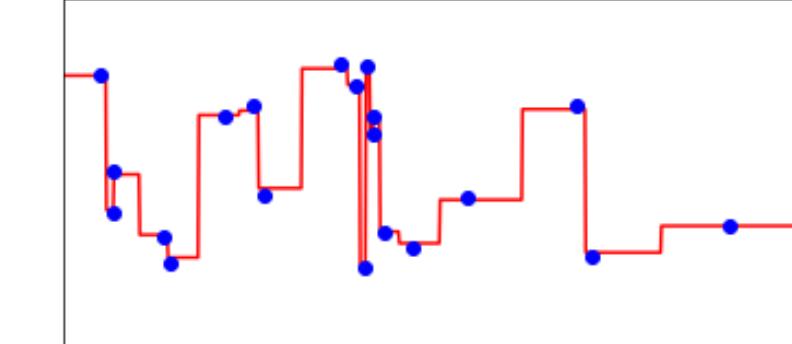
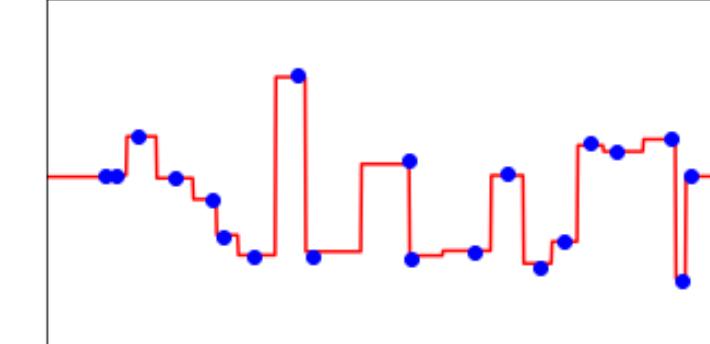
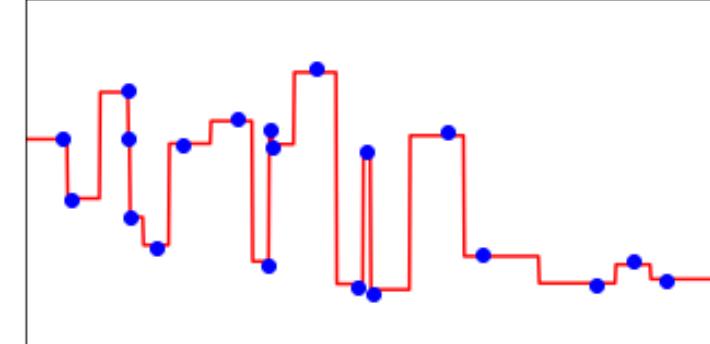
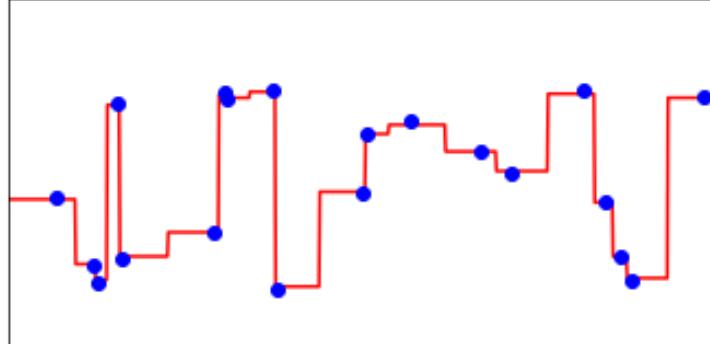
Gradient Boosted Trees



Gradient Boosted Trees

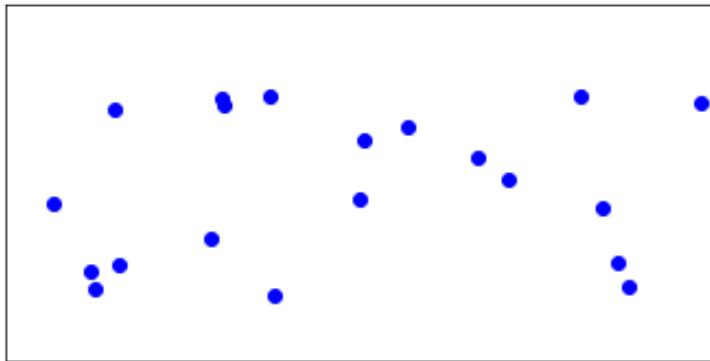


Gradient Boosted Trees

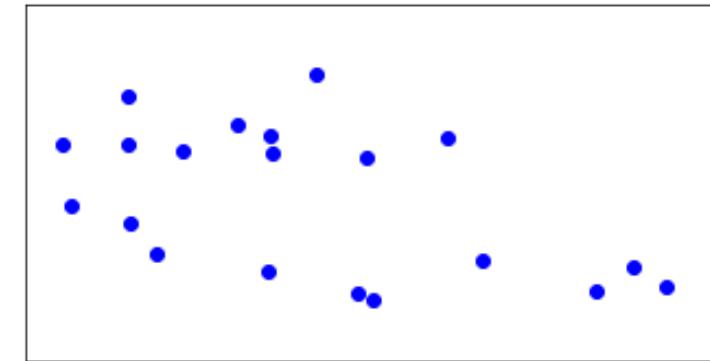


局所性をもつInterpolator

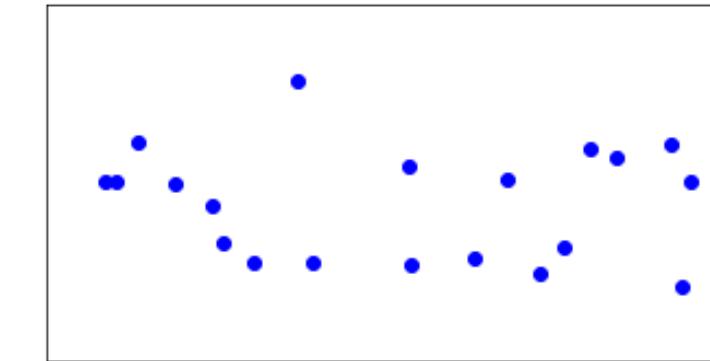
全くランダムなラベルでも訓練誤差0を達成できる（この設定ではGBDT, 1-NN, DTはほぼ同じ）



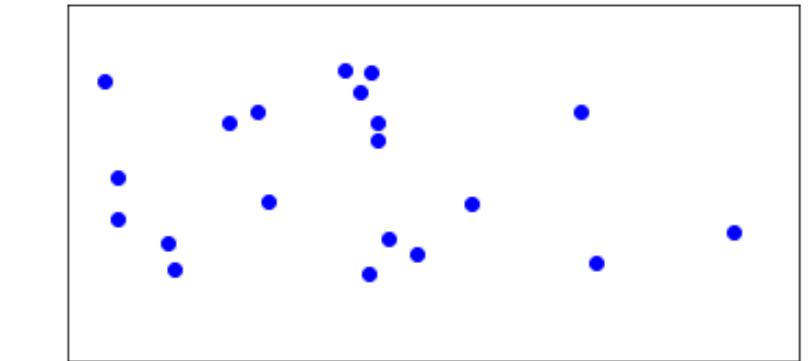
Nearest Neighbor ($k=1$)



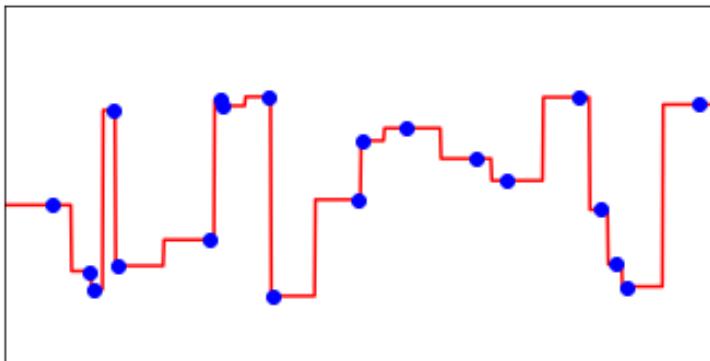
Nearest Neighbor ($k=1$)



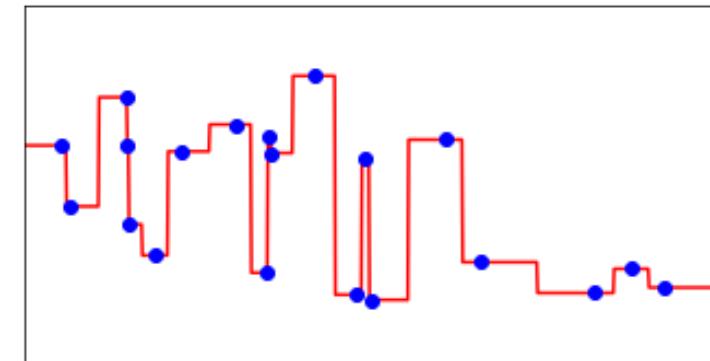
Nearest Neighbor ($k=1$)



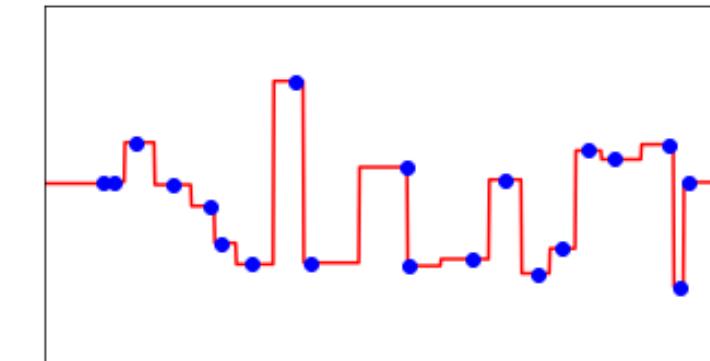
Nearest Neighbor ($k=1$)



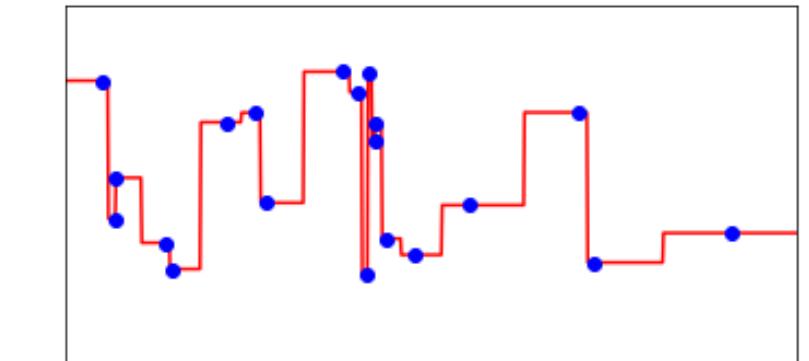
Decision Tree



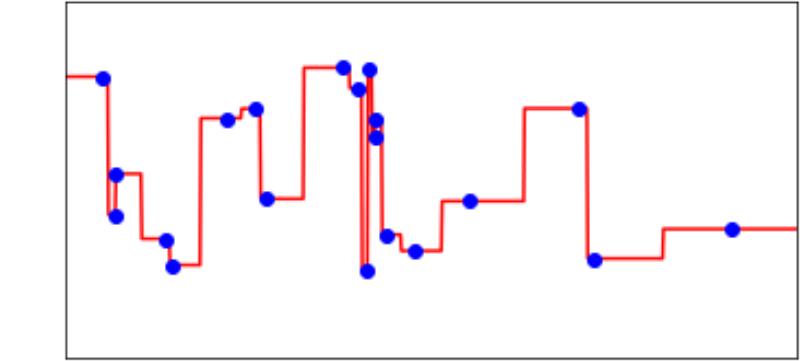
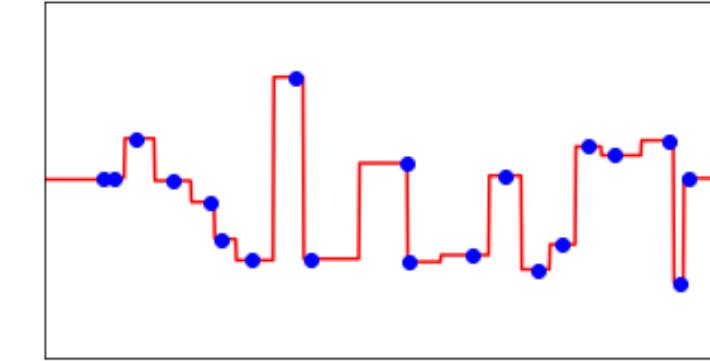
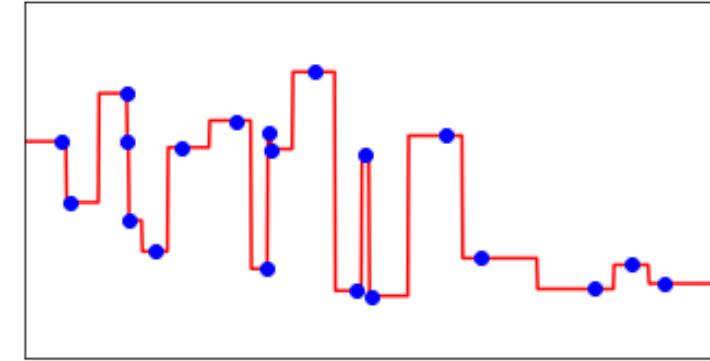
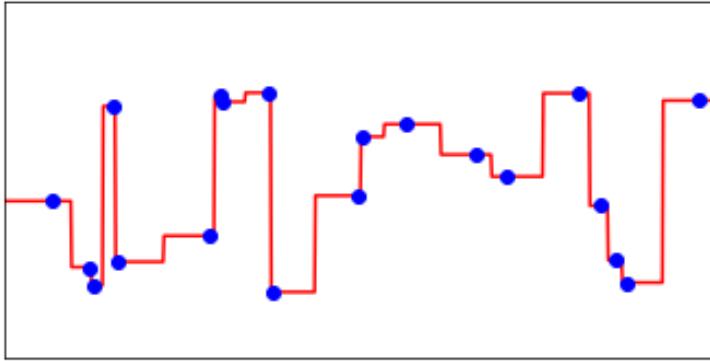
Decision Tree



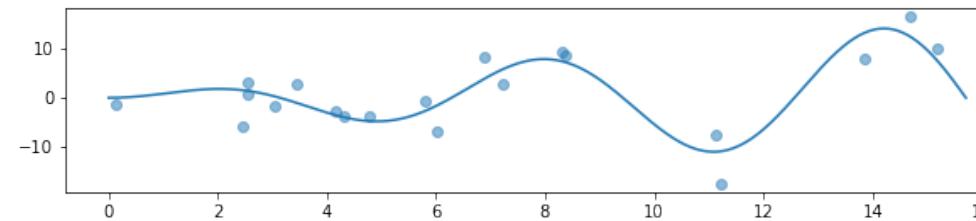
Decision Tree



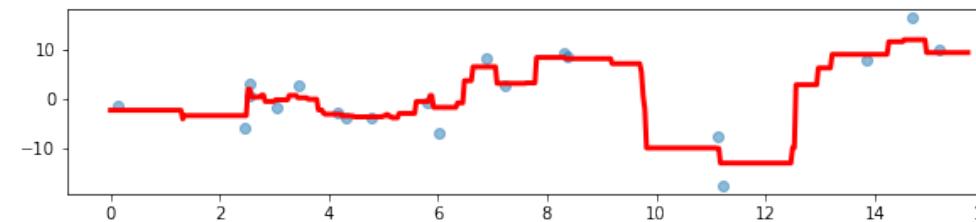
Decision Tree



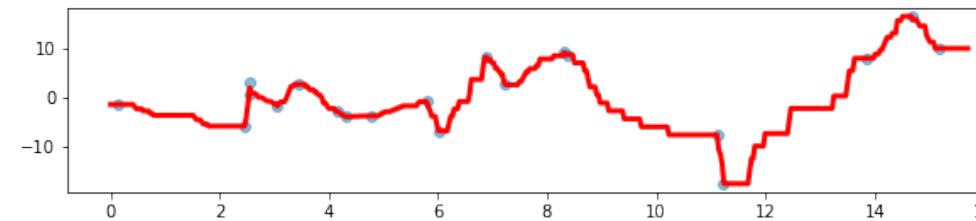
回帰が多い材料タスクでは特にExtra Treesが初手にオススメ



RandomForestRegressor(*n_estimators*=10)



ExtraTreesRegressor(*n_estimators*=10)



サンプル点間の自然な連続補間の
ような性質を持つ（高次元でも！）

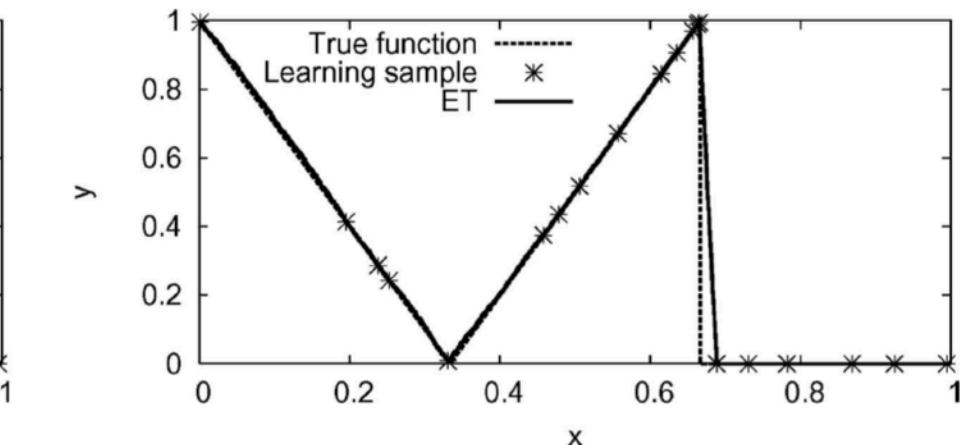
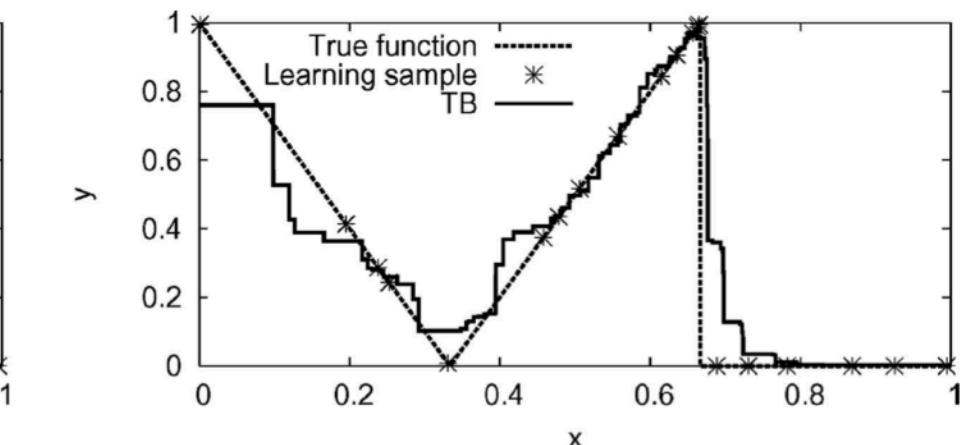
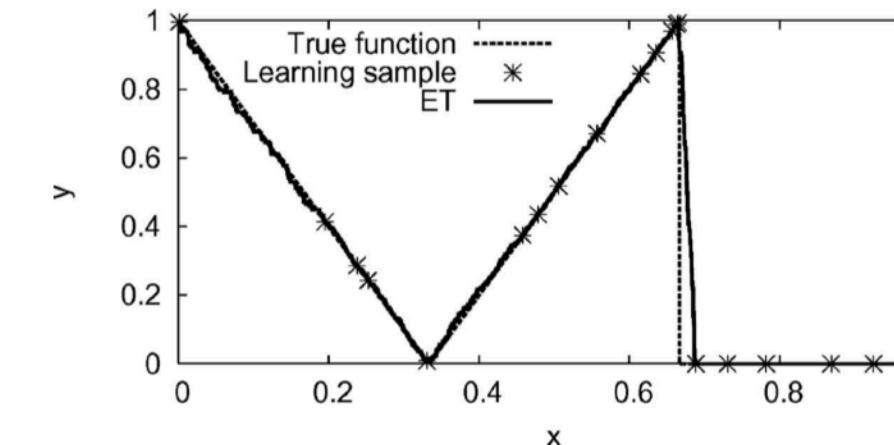
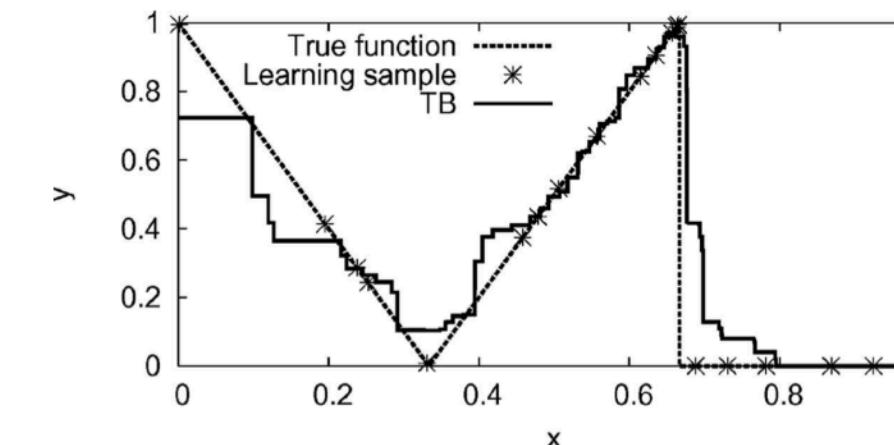


Fig. 10 Tree Bagging, and fully developed Extra-Trees ($n_{\min} = 2$) on a one-dimensional piecewise linear problem ($N = 20$). Left with $M = 100$ trees, right with $M = 1000$ trees.

Geurts, Ernst, Wehenkel, Extremely randomized trees. *Mach Learn* 63, 3–42 (2006).

<https://doi.org/10.1007/s10994-006-6226-1>

Local-averaging estimators

ヒストグラムやNearest Neighborをバカにするなれ。こうした局所平均化推定量は理論的にも非常に良い性質をもち、分布が複雑すぎてモデル化が著しく困難な現代のデータにマッチ？

Local-averaging estimators

ヒストグラムやNearest Neighborをバカにするなれ。こうした局所平均化推定量は理論的にも非常に良い性質をもち、分布が複雑すぎてモデル化が著しく困難な現代のデータにマッチ？

- **Nearest Neighbor推定量**

存在するかどうか不明であった「 (X, y) の分布がどんな分布でも一致性を持つ」夢の性質
ユニバーサル一致性をもつ推定量であることが示された(Stoneの定理, 1977)

- **Nadaraya-Watson推定量**

この議論でよく登場する単純なカーネル推定量。「**Attention Mechanism**」の最初期の例としてICML2019の「A Tutorial on Attention in Deep Learning」でも紹介された。

<https://icml.cc/Conferences/2019/ScheduleMultitrack?event=4343>

- **Histogram rules on data-dependent partitions (or data-driven histogram methods)**

決定木を含む。90年代後半に研究された(Nobel, Ann. Statist. 24(3), 1996; Lugosi & Nobel, Ann. Statist. 24(2), 1996)。詳しくは下記書籍(通称 Yellow Terror)を参照。

参考：現代のInterpolation論

ラベルノイズがある場合ですらも訓練データを完全に内挿する(訓練誤差0 or ほぼ0)方がテスト誤差も小さくなる現象はHarmless overfittingやBenign overfittingとして注目されている。(深層学習界隈の議論が中心だが、高次元では線形学習でも起こることは驚きを与えた)

参考：現代のInterpolation論

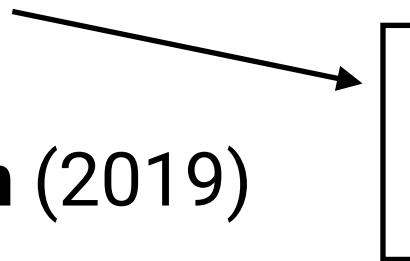
ラベルノイズがある場合ですらも訓練データを完全に内挿する(訓練誤差0 or ほぼ0)方がテスト誤差も小さくなる現象はHarmless overfittingやBenign overfittingとして注目されている。(深層学習界隈の議論が中心だが、高次元では線形学習でも起こることは驚きを与えた)

Belkin et al, **Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate** (2018)
<https://arxiv.org/abs/1806.05161>

Hastie et al, **Surprises in high-dimensional ridgeless least squares interpolation** (2020)
<https://arxiv.org/abs/1903.08560>

Bartlett et al, **Benign overfitting in linear regression** (2019)
<https://arxiv.org/abs/1906.11300>

Muthukumar et al, **Harmless interpolation of noisy data in regression** (2019)
<https://arxiv.org/abs/1903.09139>


$$\begin{array}{ll} \text{Ridgeless} & \beta = (X'X)^+ X'y \\ \text{Ridge} & \beta = (X'X + \lambda I)^{-1} X'y \end{array}$$

参考：現代のInterpolation論

ラベルノイズがある場合ですらも訓練データを完全に内挿する(訓練誤差0 or ほぼ0)方がテスト誤差も小さくなる現象はHarmless overfittingやBenign overfittingとして注目されている。(深層学習界隈の議論が中心だが、高次元では線形学習でも起こることは驚きを与えた)

*"Many modern machine learning models are **trained to achieve zero or near-zero training error** in order to obtain **near-optimal (but non-zero) test error**. This phenomenon of strong generalization performance for "overfitted" / interpolated classifiers appears to be ubiquitous in high-dimensional data, having been observed in deep networks, kernel machines, boosting and random forests. **Their performance is consistently robust even when the data contain large amounts of label noise.**"*

Weyner AJ, Olson M, Bleich J, and Mease D.

[Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers](#). *J Mach Learn Res.*

2017; 18(48): 1-33. <https://jmlr.org/papers/v18/15-240.html>

Belkin M, Hsu D, Mitra PP.

[Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate](#). *NIPS 2018*.

<https://dl.acm.org/doi/10.5555/3327144.3327157>

Belkin M.

[Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation](#).

Acta Numerica, 2021; 30: 203-248: <https://doi.org/10.1017/S0962492921000039>

最初に議論を提起した
めちゃ面白い論文

参考：深層学習も局所的な領域分割に基づく予測

経験的には深層学習は**局所性鋭敏型ハッシュ (LSH)**のようと言われてきた。



...

One way to think of a neural network is as a hashtable where the hashing function is locality-sensitive. It memorizes training inputs & targets, and is capable of successfully querying targets for test inputs that are very close to what it has already seen.

3:33 PM · Aug 9, 2018

89 Retweets 14 Quote Tweets 357 Likes

参考：深層学習も局所的な領域分割に基づく予測

経験的には深層学習は**局所性鋭敏型ハッシュ (LSH)**のようと言われてきた。



François Chollet @fchollet

One way to think of a neural network is as a hashtable where the hashing function is locality-sensitive. It memorizes training inputs & targets, and is capable of successfully querying targets for test inputs that are very close to what it has already seen.

3:33 PM · Aug 9, 2018

89 Retweets 14 Quote Tweets 357 Likes

現行の深層学習は**局所的領域分割に基づく区分線形モデル (自由節点の線形スプライン)**と見なせる。

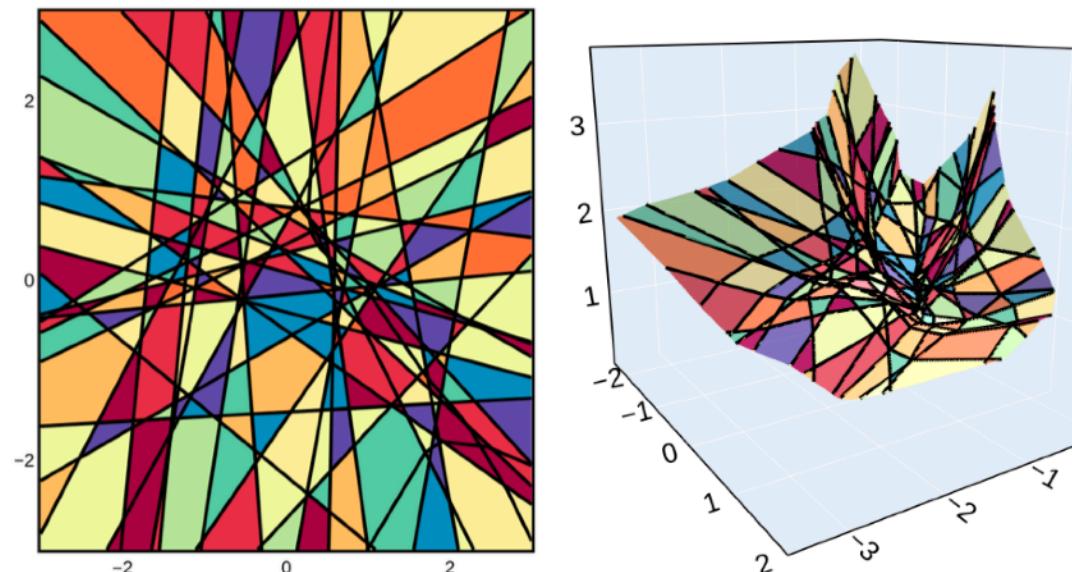
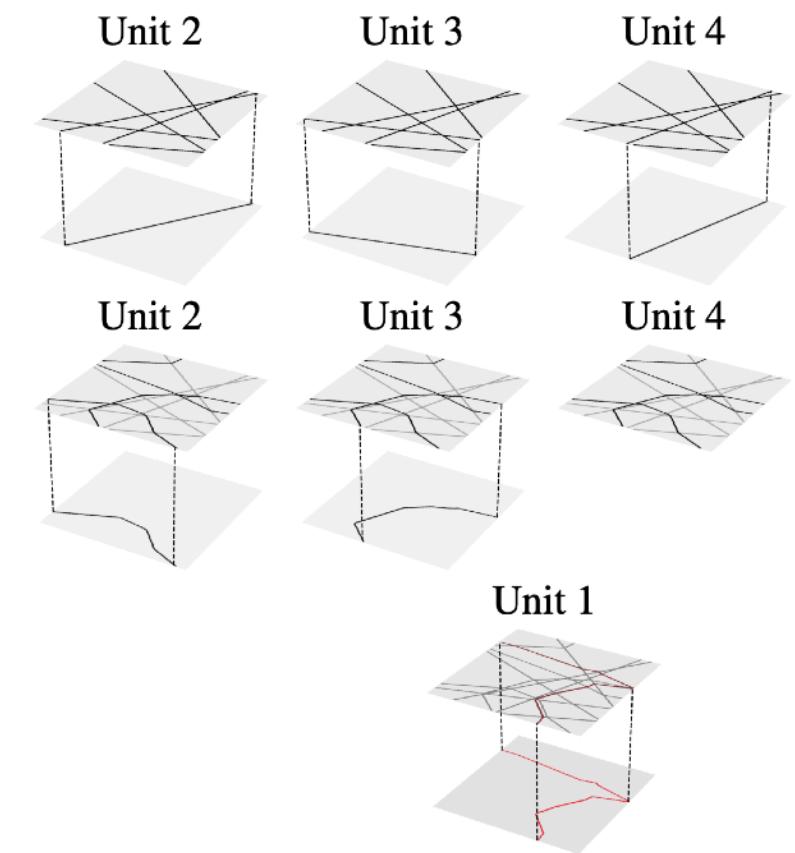


Figure 4.1 : Visual depiction of Thm. 4.1 with a (random) generator $G : \mathbb{R}^2 \mapsto \mathbb{R}^3$.



ReLU = $\max(x, 0)$ は**空間を二分割し、片側だけに着目する操作**
(= 負値をとった領域を全無視する操作)

cf. 決定木も再帰的空間二分割

参考：深層学習も局所的な領域分割に基づく予測

経験的には深層学習は**局所性鋭敏型ハッシュ (LSH)**のようと言われてきた。



François Chollet

One way to think of a neural network is as a hashtable where the hashing function is locality-sensitive. It memorizes training inputs & targets, and is capable of successfully querying targets for test inputs that are very close to what it has already seen.

3:33 PM · Aug 9, 2018

89 Retweets 14 Quote Tweets 357 Likes

現行の深層学習は**局所的領域分割に基づく区分線形モデル (自由節点の線形スプライン)**と見なせる。

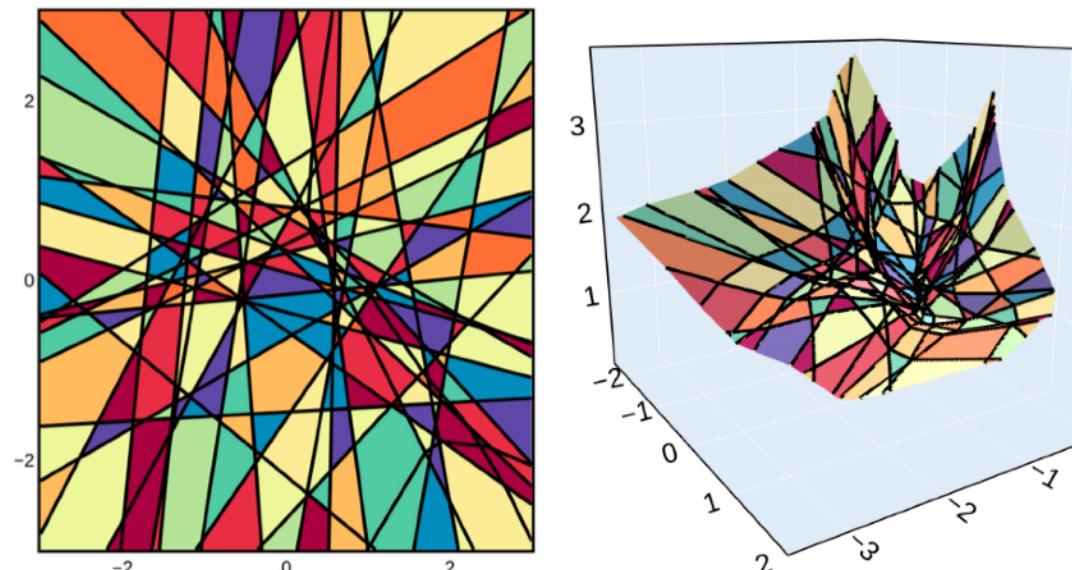
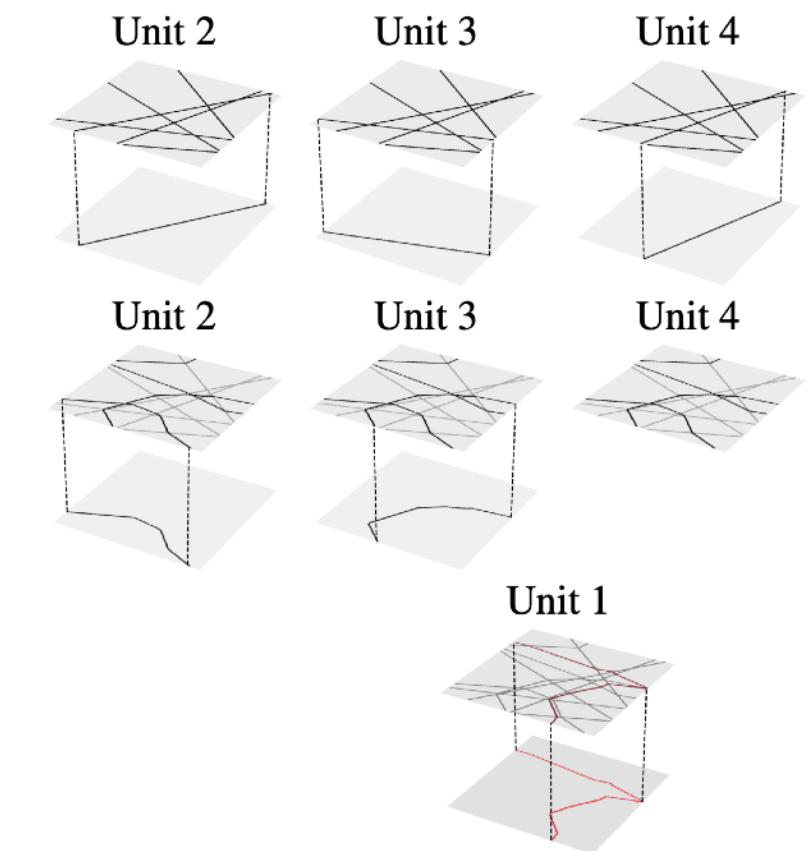


Figure 4.1 : Visual depiction of Thm. 4.1 with a (random) generator $G : \mathbb{R}^2 \mapsto \mathbb{R}^3$.

Balestrieri, Randall. "Max-Affine Splines Insights Into Deep Learning." (2021)
Diss., Rice University. <https://hdl.handle.net/1911/110439>

Balestrieri & Baraniuk. A Spline Theory of Deep Learning (ICML 2018)
<https://proceedings.mlr.press/v80/balestrieri18b.html>

Balestrieri et al., The Geometry of Deep Networks: Power Diagram Subdivision (NeurIPS 2019)
<https://arxiv.org/abs/1905.08443>



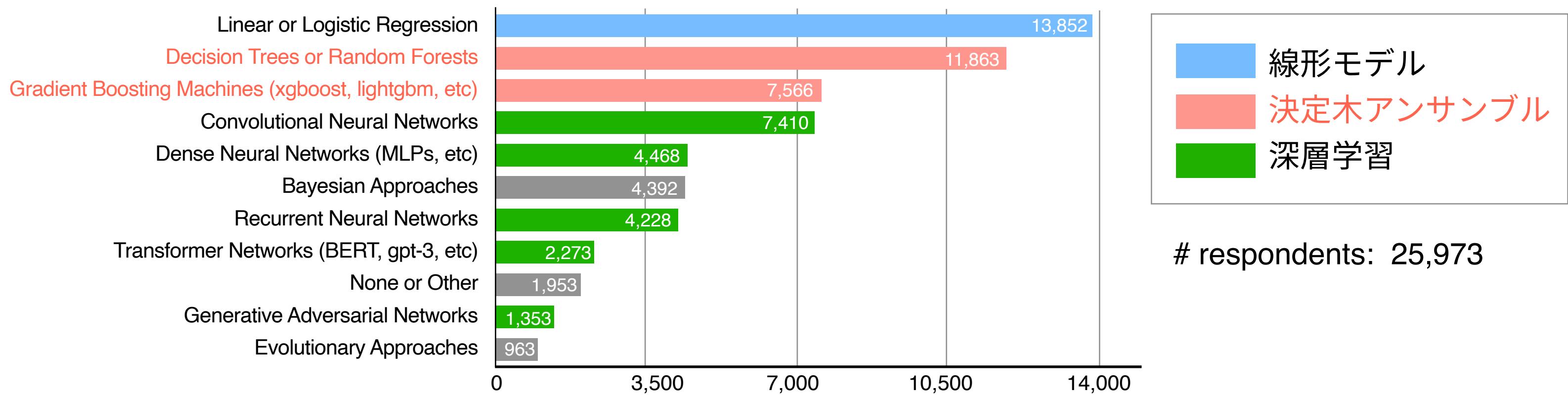
ReLU = $\max(x, 0)$ は**空間を二分割し、片側だけに着目する操作**
(= 負値をとった領域を全無視する操作)
cf. 決定木も再帰的空間二分割

決定木アンサンブルは現代のデータ科学の三種の神器の一つ

決定木アンサンブルは「実際のデータ」の解析で非常によく使われる道具
(線形モデル・深層学習と合わせて現代データサイエンスの**三種の神器の一つ?**)

State of Data Science and Machine Learning 2021 (Kaggle Survey)

Q17. Which of the following ML algorithms do you use on a regular basis?



決定木アンサンブルに基づくモデルベース探索

ExtraTreesやRandom Forestsは各領域に落ちたサンプルの分散から予測分散を自然に計算できる(勾配Boosting木の場合はQuantile回帰を使うのが王道)で自分で書くのは特に難しくはないが…

実践的にはとりあえず **scikit-optimize** の **決定木アンサンブル** を使うとカンタン！

<https://scikit-optimize.github.io/>

skopt.learning: Machine learning extensions for model-based optimization.

Machine learning extensions for model-based optimization.

User guide: See the [Learning](#) section for further details.

<code>learning.ExtraTreesRegressor([n_estimators, ...])</code>	ExtraTreesRegressor that supports conditional standard deviation.
<code>learning.GaussianProcessRegressor([kernel, ...])</code>	GaussianProcessRegressor that allows noise tunability.
<code>learning.GradientBoostingQuantileRegressor(...)</code>	Predict several quantiles with one estimator.
<code>learning.RandomForestRegressor(...)</code>	RandomForestRegressor that supports conditional std computation.

`!pip install scikit-optimize`

```
from sklearn.datasets import load_diabetes
import skopt
X, y = load_diabetes(return_X_y=True)
model = skopt.learning.ExtraTreesRegressor()
model.fit(X, y)
X_test = np.random.rand(1, X.shape[1])
y_pred, y_std = model.predict(X_test, return_std=True)
```

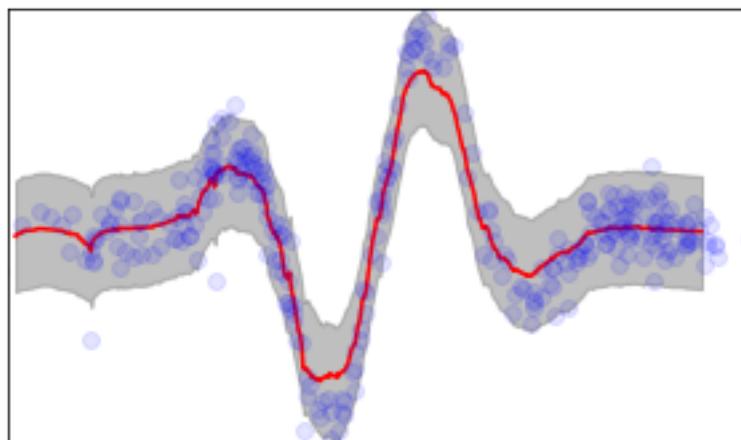
近年のConformal Predictionの発展

Conformal Predictionを使っても良い

<https://github.com/scikit-learn-contrib/MAPIE>

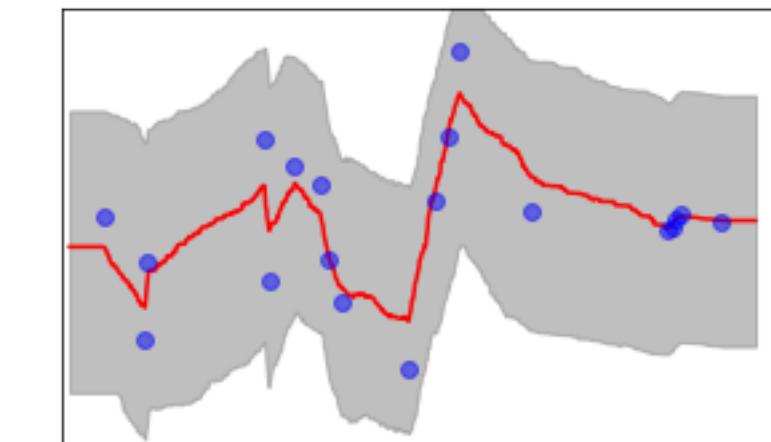
`regressor = ExtraTreesRegressor(max_leaf_nodes=32, bootstrap=True)`

MAPIE



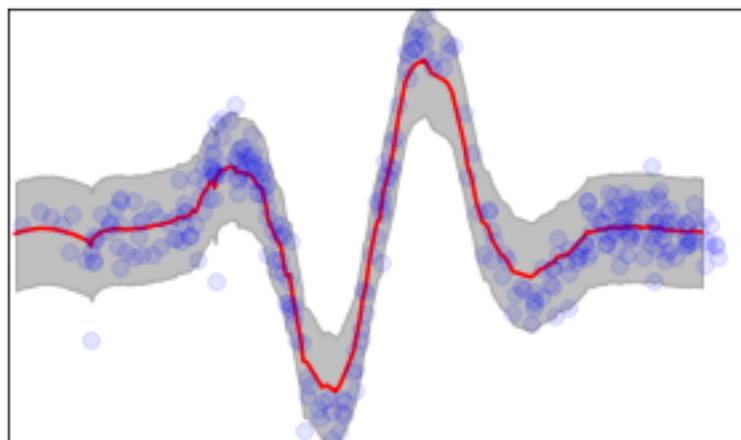
`MapieRegressor(regressor,
method="plus", cv=-1)`

*Jackknife+
95% prediction intervals*



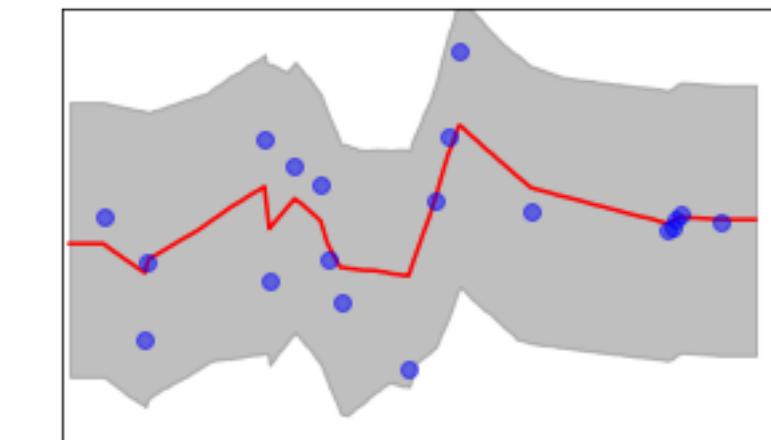
`MapieRegressor(regressor,
method="plus", cv=-1)`

*Jackknife+
90% prediction intervals*



`MapieRegressor(regressor,
method="plus",
cv=Subsample(n_resamplings=50))`

*Jackknife+ after bootstrap
95% prediction intervals*



`MapieRegressor(regressor,
method="plus",
cv=Subsample(n_resamplings=50))`

*Jackknife+ after bootstrap
90% prediction intervals*

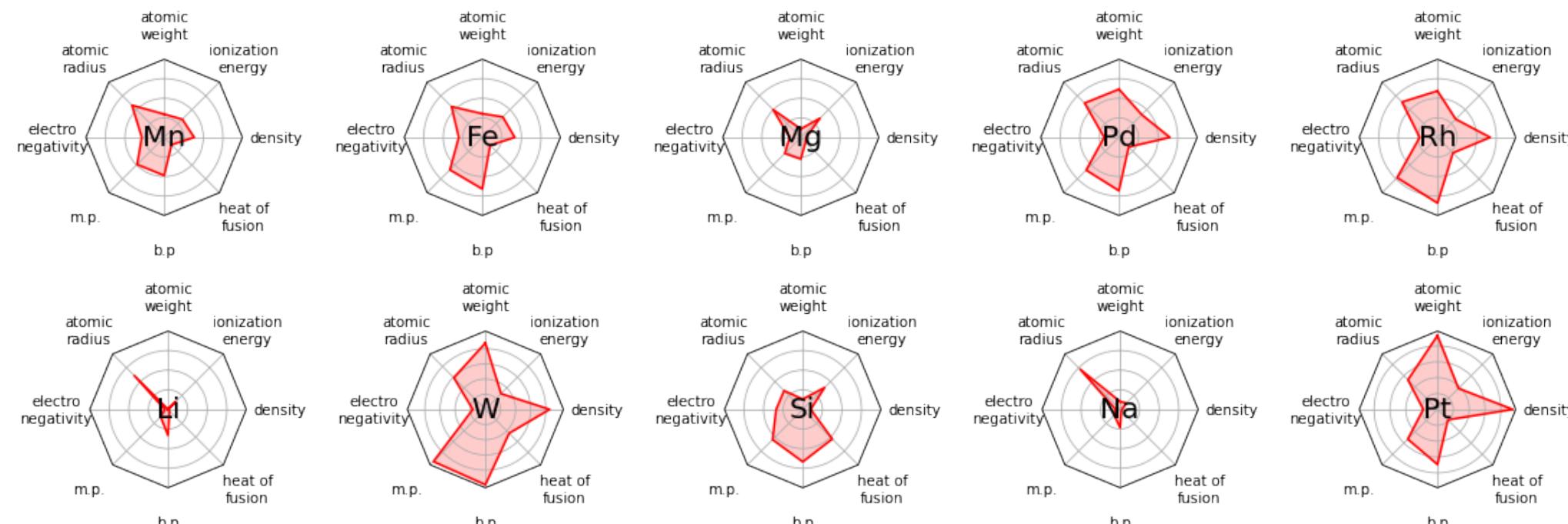
保守的すぎて訓練データの微改善しかできないのでは？

「訓練データの微改善」を反復するのが機械発見の本質 + ランダム探索と併用

+ 例えば、対象の変数表現の粒度で探索のアグレッシブさのコントロールが可能

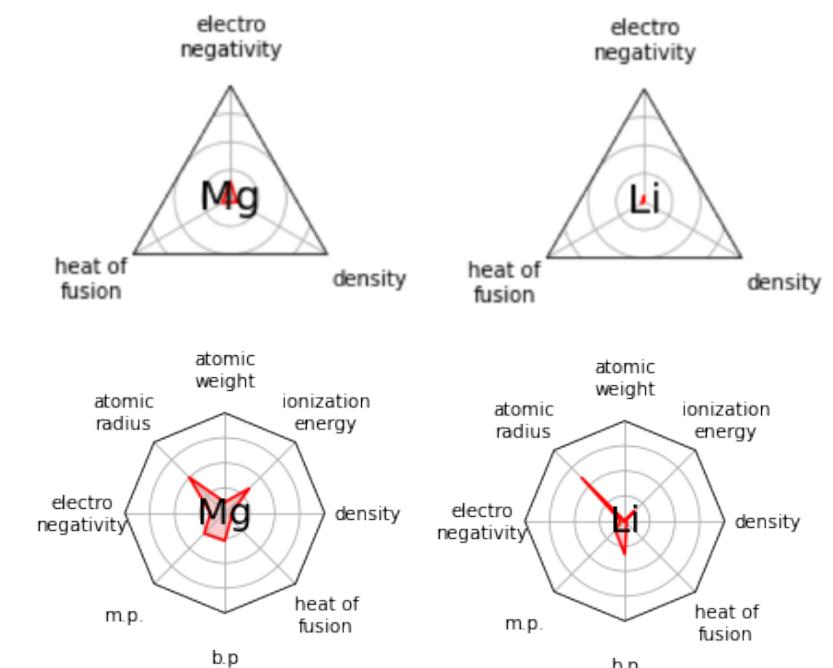
元素をカテゴリカル値で扱うのではなく 元素記述子表現で扱うことで訓練データで全く使われていない元素を探索に取り入れることが可能（実際にそのようなヒットが見つかっている）

Wang et al., Accelerated discovery of multi-elemental reverse water-gas shift catalysts using extrapolative machine learning approach. (2022) <https://doi.org/10.26434/chemrxiv-2022-695rj>



SWED-3

SWED-8



今日のたった3つの話

お題「材料科学における機械学習」に沿って、この幻想が打ち砕かれてゆく過程で、機械学習研究者として得た教訓と最新知見をもとに、次の3つの主張をしてみたいと思います…

1. 「機械学習」と「材料科学」はゴールが根本的に食い違っていて、機械学習とは全く設定が異なる「**機械発見**」が求められている。
2. 仮説フリーの探索では、**決定木アンサンブルによる探索**がoff-the-shelfかつ非常に強力なベースラインになる。
3. それ以上を求めるなら、**仮説フリーではいられない。**

今日の主張③

それ以上を求めるなら、**仮説フリーではいられない。**

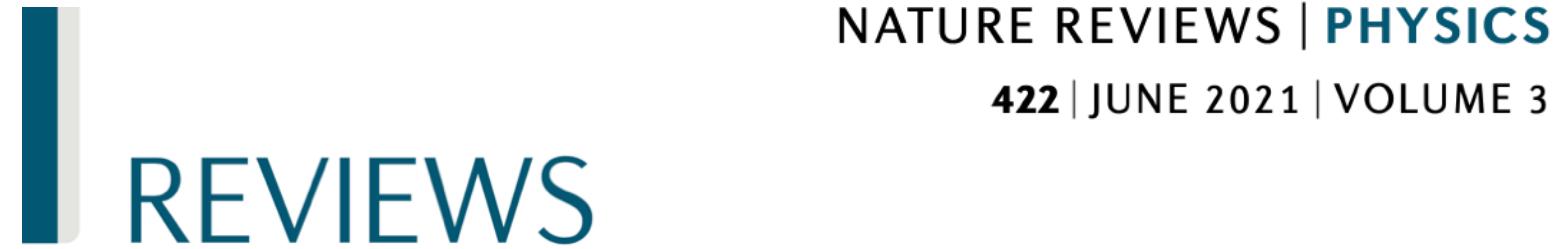
今日の主張③

それ以上を求めるなら、**仮説フリーではいられない。**

- 見本点が足りてない状況では同程度にCV精度が良い関数は**無数に存在**(族をなす)
- CV精度が良い関数であれば「**どれでも良いわけではない**」
→ 関数が少なくとも**満たして欲しい性質や条件がある**(しばしば暗黙的に)
- 教師あり学習では見本点(X_i, y_i)を $y_i \approx \hat{y}_i = f(X_i)$ と再現する関数fを一つ選択するが
それがその要件を満たす保証も対象現象を近似している保証も「**どこにもない**」
- どのが良いかをデータから決められない以上、**近似しようとしている対象現象に**
関する事前知識・理論・モデル・仮説を使って「**良い関数**」を選択する
→ **帰納と演繹(数理モデル)の融合** (ただし同時に汎用性を諦めることを意味する)

例：Physics-informed ML, Geometric ML, Causal ML

<https://doi.org/10.1038/s42254-021-00314-5>



NATURE REVIEWS | PHYSICS

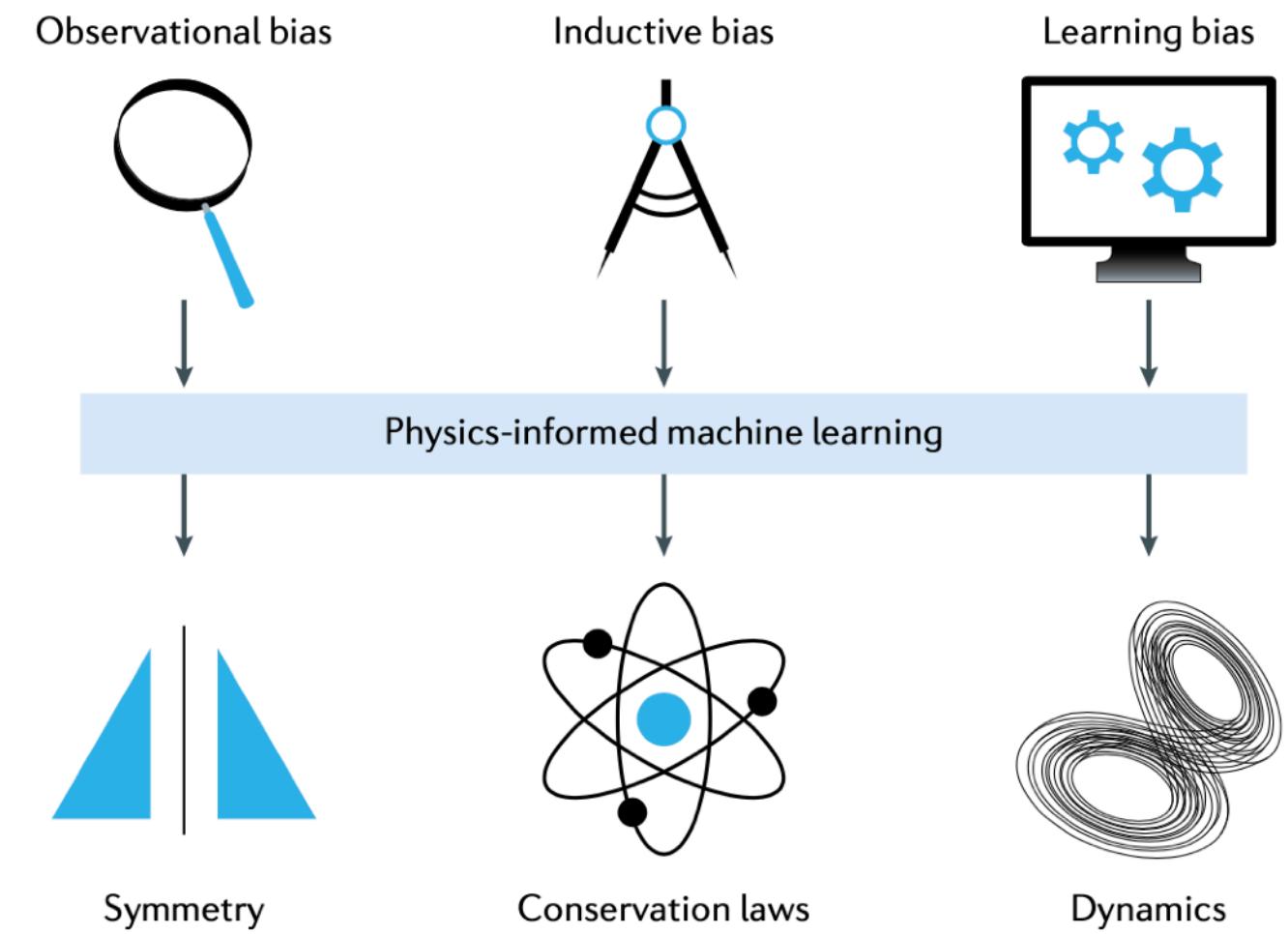
422 | JUNE 2021 | VOLUME 3

REVIEWS

Physics-informed machine learning

George Em Karniadakis^{1,2}✉, Ioannis G. Kevrekidis^{3,4}, Lu Lu¹, Paris Perdikaris⁶, Sifan Wang⁷ and Liu Yang¹

Abstract | Despite great progress in simulating multiphysics problems using the numerical discretization of partial differential equations (PDEs), one still cannot seamlessly incorporate noisy data into existing algorithms, mesh generation remains complex, and high-dimensional problems governed by parameterized PDEs cannot be tackled. Moreover, solving inverse problems with hidden physics is often prohibitively expensive and requires different formulations and elaborate computer codes. Machine learning has emerged as a promising alternative, but training deep neural networks requires big data, not always available for scientific problems. Instead, such networks can be trained from additional information obtained by enforcing the physical laws (for example, at random points in the continuous space-time domain). Such physics-informed learning integrates (noisy) data and mathematical models, and implements them through neural networks or other kernel-based regression networks. Moreover, it may be possible to design specialized network architectures that automatically satisfy some of the physical invariants for better accuracy, faster training and improved generalization. Here, we review some of the prevailing trends in embedding physics into machine learning, present some of the current capabilities and limitations and discuss diverse applications of physics-informed learning both for forward and inverse problems, including discovering hidden physics and tackling high-dimensional problems.



帰納(統計的機械学習)と演繹(数理モデル)の融合

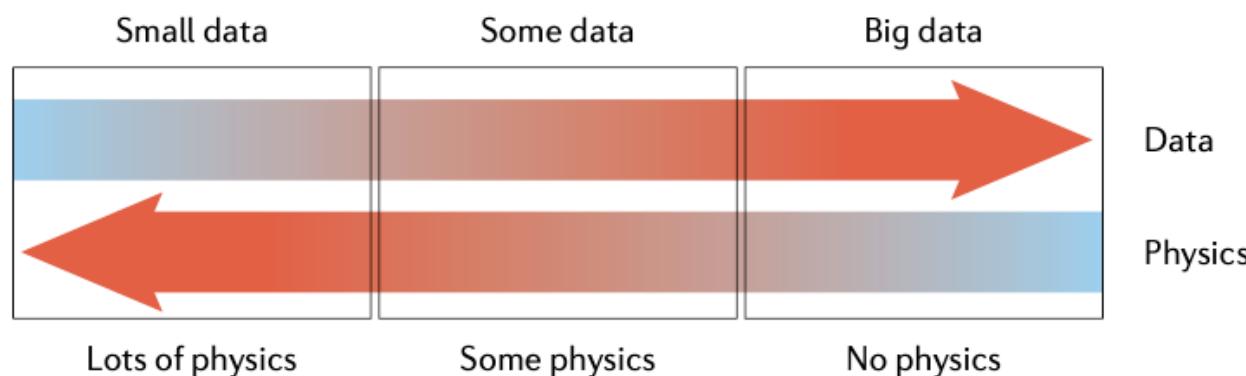
- ・シミュレーション(理論計算)がそれなりの忠実度をもつ対象の場合、シミュレーションモデルの一部内部パラメタを微分可能にしデータに同化させるのが基本
→ この場合、解はバイ・デザインで実行可能解を逸脱しない

帰納(統計的機械学習)と演繹(数理モデル)の融合

- ・ シミュレーション(理論計算)がそれなりの忠実度をもつ対象の場合、シミュレーションモデルの一部内部パラメタを微分可能にしデータに同化させるのが基本
→ この場合、解はバイ・デザインで実行可能解を逸脱しない
- ・ もう一つは機械学習をベースにおくが、汎用的な関数モデルを使うのではなく、バイ・デザインで「条件を満たす関数の範囲でのみ」フィッティングを行う方法
(不变性や同変性などのSymmetryをモデル構造に課すGeometric MLなど)

帰納(統計的機械学習)と演繹(数理モデル)の融合

- ・ シミュレーション(理論計算)がそれなりの忠実度をもつ対象の場合、シミュレーションモデルの一部内部パラメタを微分可能にしデータに同化させるのが基本
→ この場合、解はバイ・デザインで**実行可能解を逸脱しない**
- ・ もう一つは機械学習をベースにおくが、汎用的な関数モデルを使うのではなく、バイ・デザインで「**条件を満たす関数の範囲でのみ**」フィッティングを行う方法(不变性や同変性などのSymmetryをモデル構造に課すGeometric MLなど)



帰納と演繹をハイブリッドさせる場合には
設計上のバランスの自由度がある！

ガチガチの演繹をちょっと帰納で緩和する
↔ ユルユルな帰納をちょっと演繹で制約する

失われた因果を求めて：因果推論(因果分析)と因果探索



- ナイルズは、因果分析の目的を、**XがYの原因だと証明すること**、あるいは**Yの原因を一から見つけ出すこと**だと考えていた。この**誤った考え方**は、今でも多くの人に見受けられる。だがそれは因果分析ではなく、「因果探索」が扱う問題だ。

失われた因果を求めて：因果推論(因果分析)と因果探索



- ナイルズは、因果分析の目的を、**XがYの原因だと証明すること**、あるいは**Yの原因を一から見つけ出すこと**だと考えていた。この**誤った考え方**は、今でも多くの人に見受けられる。だがそれは因果分析ではなく、「因果探索」が扱う問題だ。
- 因果分析は、**ただデータだけがあれば成り立つものではない**。因果分析は、データが作られるプロセスをある程度以上、理解しないければできないことだ。つまり、**データだけではわからな**いことをはじめからある程度、知っていなくてはいけない。

失われた因果を求めて：因果推論(因果分析)と因果探索



- ナイルズは、因果分析の目的を、**XがYの原因だと証明すること**、あるいは**Yの原因を一から見つけ出すこと**だと考えていた。この**誤った考え方**は、今でも多くの人に見受けられる。だがそれは因果分析ではなく、「因果探索」が扱う問題だ。
- 因果分析は、**ただデータだけがあれば成り立つものではない**。因果分析は、データが作られるプロセスをある程度以上、理解しないければできないことだ。つまり、**データだけではわからぬことをはじめからある程度、知っていなくてはいけない**。
- 因果分析は、相関関係だけではなく、主流の統計学を構成する他のほどのような道具とも異質である。それは、因果分析が**その使い手に主観的な関与を要求する**からだ。

データの変容によって統計学は根本的変革を迫られている

1. Two Cultures (統計学 vs 機械学習)：予測と解釈は分けられるのか？

- ▶ 現況を彷彿とさせるData Modeling (統計学) vs. Algorithmic Modeling (機械学習)の対比
L. Breiman “Statistical Modeling: The Two Cultures” (2001) + CoxやEfronらのコメント
- ▶ 20周年を記念し振り返るObservational Studies誌の特集企画 (2021) + 解釈の問題の再燃
Breiman's main point is: If you want prediction, do prediction for its own sake and
forget about the illusion of representing nature. (Judea Pearl)

データの変容によって統計学は根本的変革を迫られている

1. Two Cultures (統計学 vs 機械学習)：予測と解釈は分けられるのか？

- ▶ 現況を彷彿とさせるData Modeling (統計学) vs. Algorithmic Modeling (機械学習)の対比
L. Breiman “Statistical Modeling: The Two Cultures” (2001) + CoxやEfronらのコメント
- ▶ 20周年を記念し振り返るObservational Studies誌の特集企画 (2021) + 解釈の問題の再燃
Breiman's main point is: If you want prediction, do prediction for its own sake and forget about the illusion of representing nature. (Judea Pearl)

2. もう「p値」は引退させよう！(誤用・濫用やHARKingの問題、再現性の危機)

- ▶ 統計学は道を照らす明かりというより酔っ払いのための支柱 (A. Lang, 1910)
- ▶ 2019年 Nature “Retire statistical significance” “Don't say statistically significant”
アメリカ統計学会(ASA)での宣言(2016)・学会誌特集(2019)
“我々は統計的有意性の概念全体を放棄することを求める” (800人以上の専門家が署名)

科学と仮説：科学的理解vs科学的発見

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」



科学と仮説：科学的理解 vs 科学的発見

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」



- 「予測」ではなく「理解」はあくまで人間側の認知の問題であり、科学的理解を求めるなら (=因果の理解を求めるなら)、**基本的に仮説フリーではいられない。**
- “*Theory-driven models can be wrong. But data-driven models cannot be wrong or right. Data-driven are not trying to describe an underlying reality.*”
David Hand, KDD2018 (Keynote Talk)
http://videolectures.net/kdd2018_hand_data_science/
- 一方、「発見」は探索空間が明確な場合は高次元探索で解ける潜在性がある！
(私が機械理解や解釈的機械学習より機械発見に魅力を感じている理由)

機械発見：「演繹しかない世界」の終わりの始まり…？

「材料科学・物質科学×機械学習」の分野にとどまらず、演繹的方法が唯一の方法であったあらゆる分野で、「予測」や「学習」の組込みによる発見が始まっている！

ただし下記はいずれも「問題が記述可能かつclosed」で「データ/試行が低コスト」

- プランニング・論理推論・組合せ最適化 (NP-hard問題の求解)
Schrittwieser et al. *Mastering Atari, Go, chess and shogi by planning with a learned model.* (“MuZero”, Nature, 2020)
- 純粹数学
Davies et al. *Advancing mathematics by guiding human intuition with AI.* (Nature, 2021)
- アルゴリズム設計
Fawzi et al. *Discovering faster matrix multiplication algorithms with reinforcement learning.* (“AlphaTensor”, Nature, 2022)

現時点での私的ベストプラクティス

系も探索候補も絞られていて低次元のパラメタ最適化だけしたい実験計画法的状況
→ ガウス過程回帰など連続モデルに基づくベイズ最適化、伝統的な応答局面法など

系も探索候補も絞られているが振りたいパラメタは多数(高次元での探索)

系は絞られているが探索候補は絞りたくない(広範囲の探索)

系すら絞られておらず手元のデータから「データなし」よりマシな示唆を得たい

→ 決定木アンサンブルに基づくモデルベース探索

何らかの理由でデータが好きなだけ(コストをかけずに)取れる

→ 深層学習に基づくモデルベース探索や遺伝的アルゴリズムによる探索

対象について明確に記述可能な理論・仮説があるか高忠実度をもつ計算系がある

→ 帰納と演繹の融合 (Physics-informed ML, Causal ML, Geometric ML, ...)

まとめ：今日のたった3つの話

スライドPDF → <https://itakigawa.page.link/IBISML2022taki>

1. 「機械学習」と「材料科学」はゴールが根本的に食い違っていて、機械学習とは全く設定が異なる **「機械発見」** が求められている。
→ 訓練データの計画、テストデータの計画、UQ、高次元探索、帰納と演繹の融合、…
2. 仮説フリーの探索では、**決定木アンサンブルによる探索**がoff-the-shelfかつ非常に強力なベースラインになる。
→ 高次元探索の困難性(データ領域外の挙動が支配的)、次元の呪い、Underspecification
3. それ以上を求めるなら、**仮説フリーではない**。
→ 帰納と演繹の融合 (Physics-informed ML, Causal ML, Geometric ML)
+ 演繹問題を高次元探索として解く (Pure Math, Pure CS, 組合せ論, 論理推論/記号処理)