

統計的機械学習（応用計量分析2）第1回

2026年度

本講義について

最新の情報はCLASSを参照してください

- 統計的機械学習（応用計量分析2）（9987F10）
- 担当教員：谷本 啓
 - 非常勤講師 / 株式会社Preferred Networks リサーチャー
 - 連絡先：a31607@rs.tus.ac.jp
 - 評価方法：到達度評価（6～7割）
＋小テスト・レポート等（3～4割）
- 最新情報はCLASSを参照のこと

データに基づく意思決定のための機械学習を用いた因果推論+ α

因果推論とは

介入

・・・意思決定向け分析に用いる

例) 広告すれば売上がいくら上がる？

←本講義の主題

もしこう { した } らどう { なる } かを推量する分析技術
 { していた } { なっていた }

反実仮想

・・・(ほぼ) 振り返りの分析に用いる

例) この女性がもし男性だったら

面接で採用されていたか？

世の中にたえて桜のなかりせば

春の心はのどけからまし

在原業平



『因果機械学習入門』 ドラフト＋補助参考書

- 谷本 『因果機械学習入門 — 意思決定のための因果推論と機械学習』
ドラフト
 - https://a-tanimoto.com/pdfs/causalMLbook_r22.pdf
 - 誤りを見つけて報告してくれた人は謝辞等にて感謝します
- 補助参考書
 - 各トピックにさらに興味が出た場合には下記の本を勧めます
 - ギルボア 『不確実性下の意思決定理論』
 - パール 『統計的因果推論 モデル・推論・推測』
 - モーガン&ウィンシップ 『反事実と因果推論』

内容（予定）

進捗と内容は様子を見ながら調整します

- 1. ガイダンス・因果推論と機械学習の概論
- 2. 意思決定理論（期待効用理論）の復習、因果推論との関係
- 3. 潜在結果モデルに基づく因果推論の枠組み
- 4. 平均因果効果の推定法
- 5. 条件付き平均因果効果（CATE）の推定法
1/2：メタ学習器、二重機械学習
- 6. CATEの推定法 2/2：
決定木と決定森、深層学習に基づく方法
- 7. 構造的因果モデルとバックドア基準
- 8. 因果探索
- 9. 発展的な因果推論手法 1/2：
フロントドア調整、操作変数法
- 10. 発展的な因果推論手法 2/2：
代理変数法、回帰不連続デザイン
- 11. 発展的な意思決定理論
- 12. 強化学習
- 13. オフライン強化学習
- 14. バンディット
- 15. 到達度評価

AI（LLM）の使用について

最終責任は人間

- 本講義の内容についてAIと相談することは自由です
- ただし、最終的な責任は人間に帰属します
- 小テスト、中間課題はAIを使用できますが、最終的に自分自身で修得する目的に適う範囲での使用を勧めます
 - 身についているかは到達度評価で確認します

因果推論と機械学習

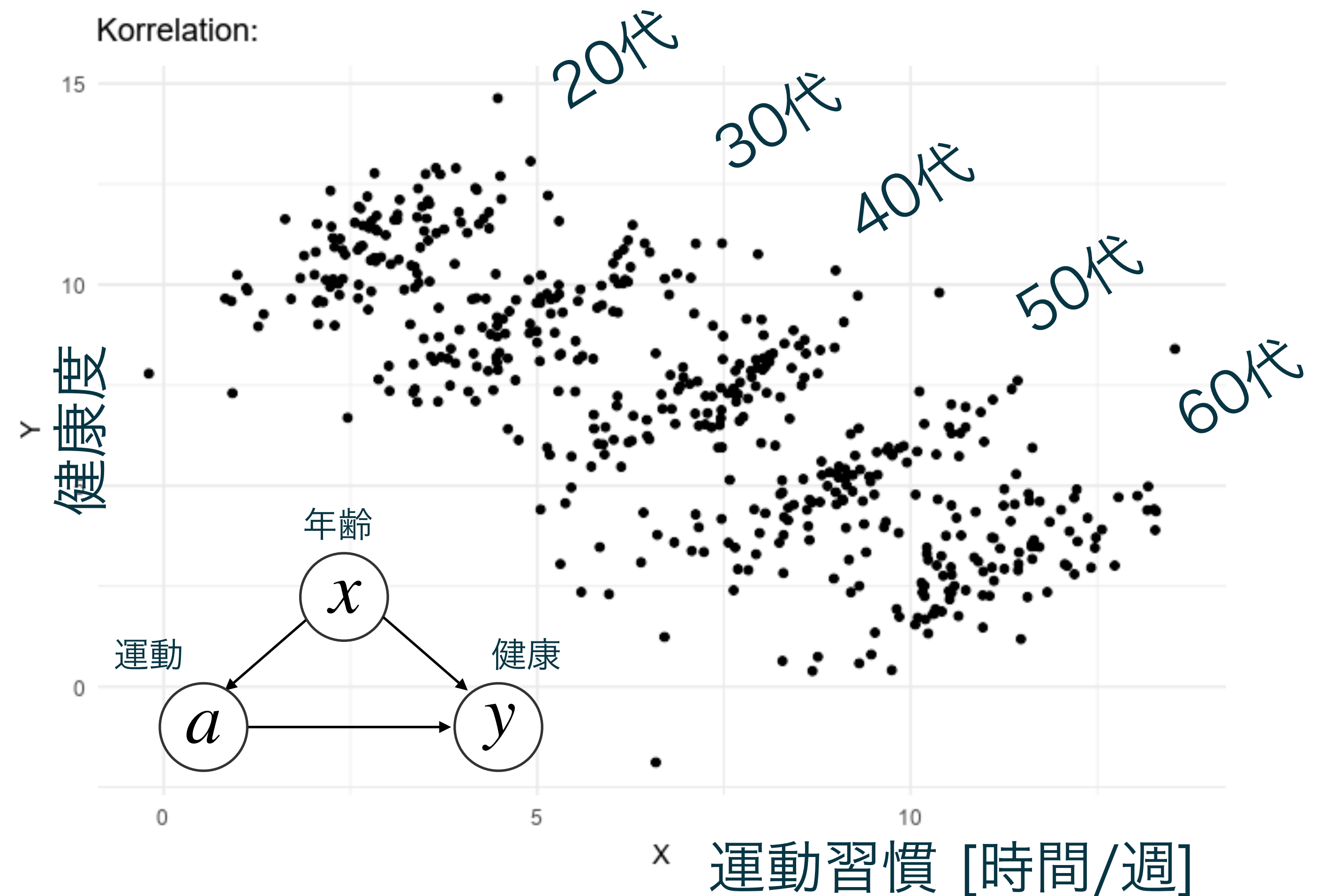
- 1. ガイダンス・因果推論と機械学習の概論
 - 因果推論の動機
 - 教師あり学習問題と一般のデータに基づく意思決定問題の違い
 - 教師あり機械学習の概説

因果推論と意思決定

動機：因果と相関の違い

シンプソンのパラドックス：データの見方で結論が変わる

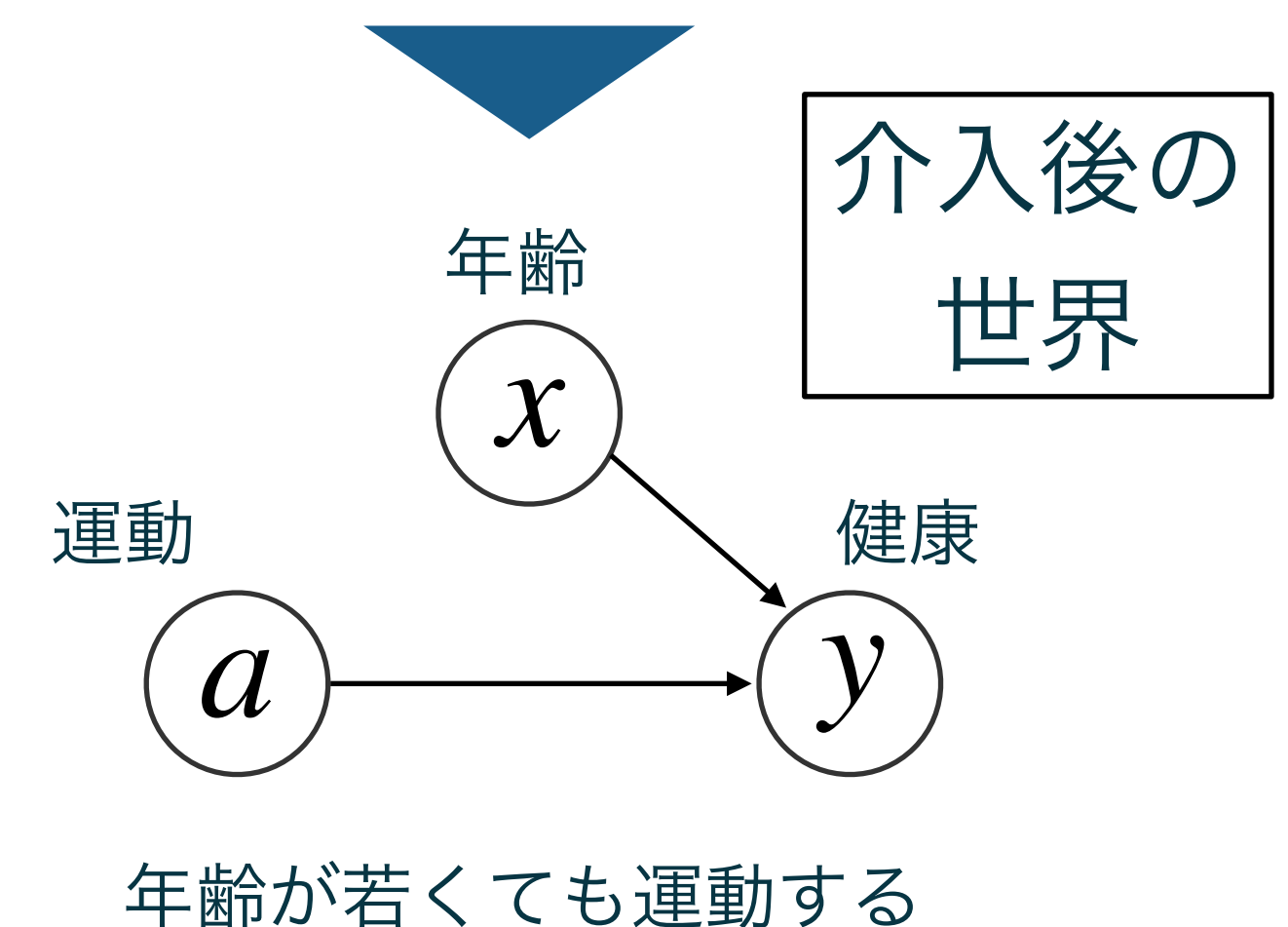
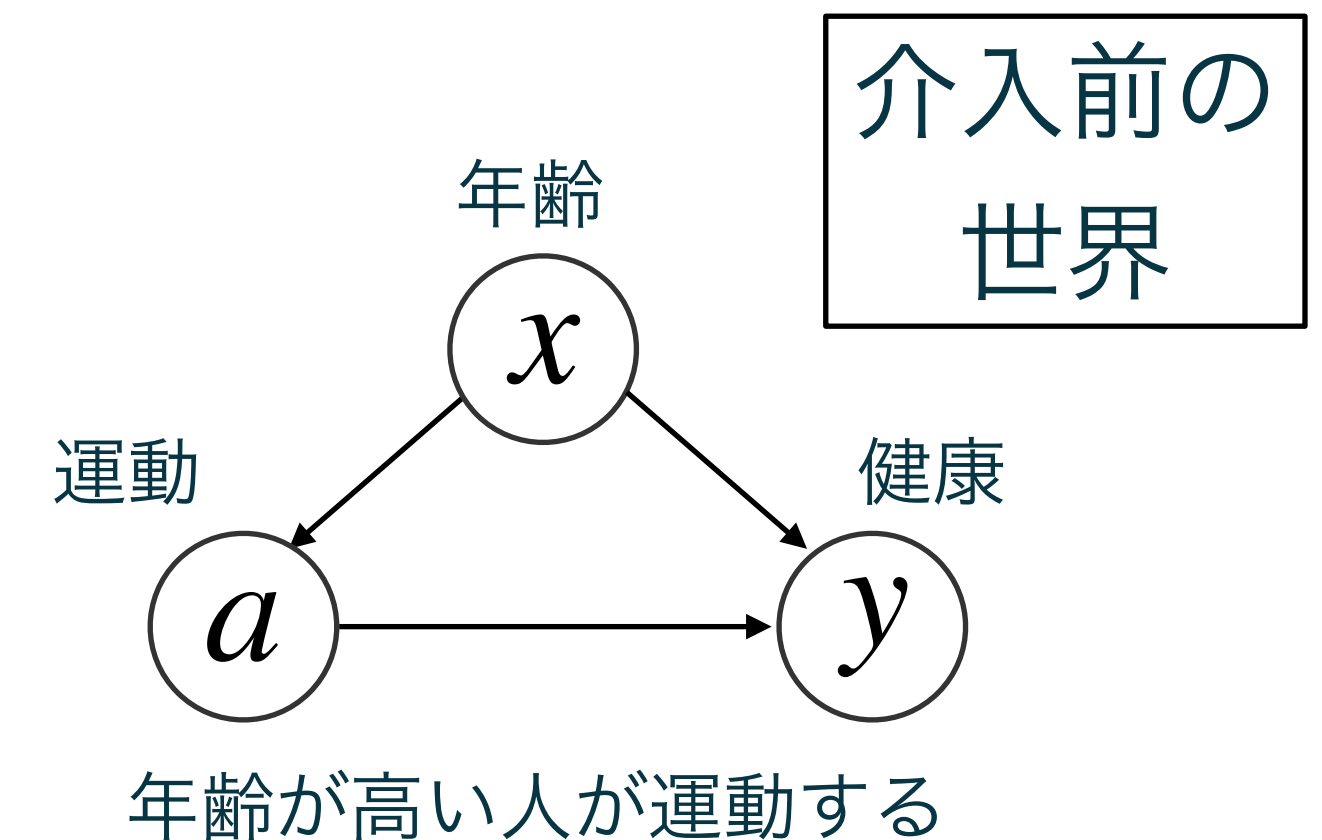
- どちらが正しい？
 - 全体で見ると運動する人ほど不健康
 - 年代別に見ると運動する人ほど健康
- →相関としてはどちらも正しい（全体の相関が層別ではどの層でも逆になる＝シンプソンのパラドックス）
- しかし
「運動する人ほど健康でない」
ことは
「運動すると健康でなくなる」
ことを意味しない（年齢が交絡因子なら因果関係を表すのは年代別の見方）
- データを何に使うのか
（意思決定）をすることが重要



意思決定に資するデータ分析のために

意思決定の仕方を変えても不変の規則性＝因果性

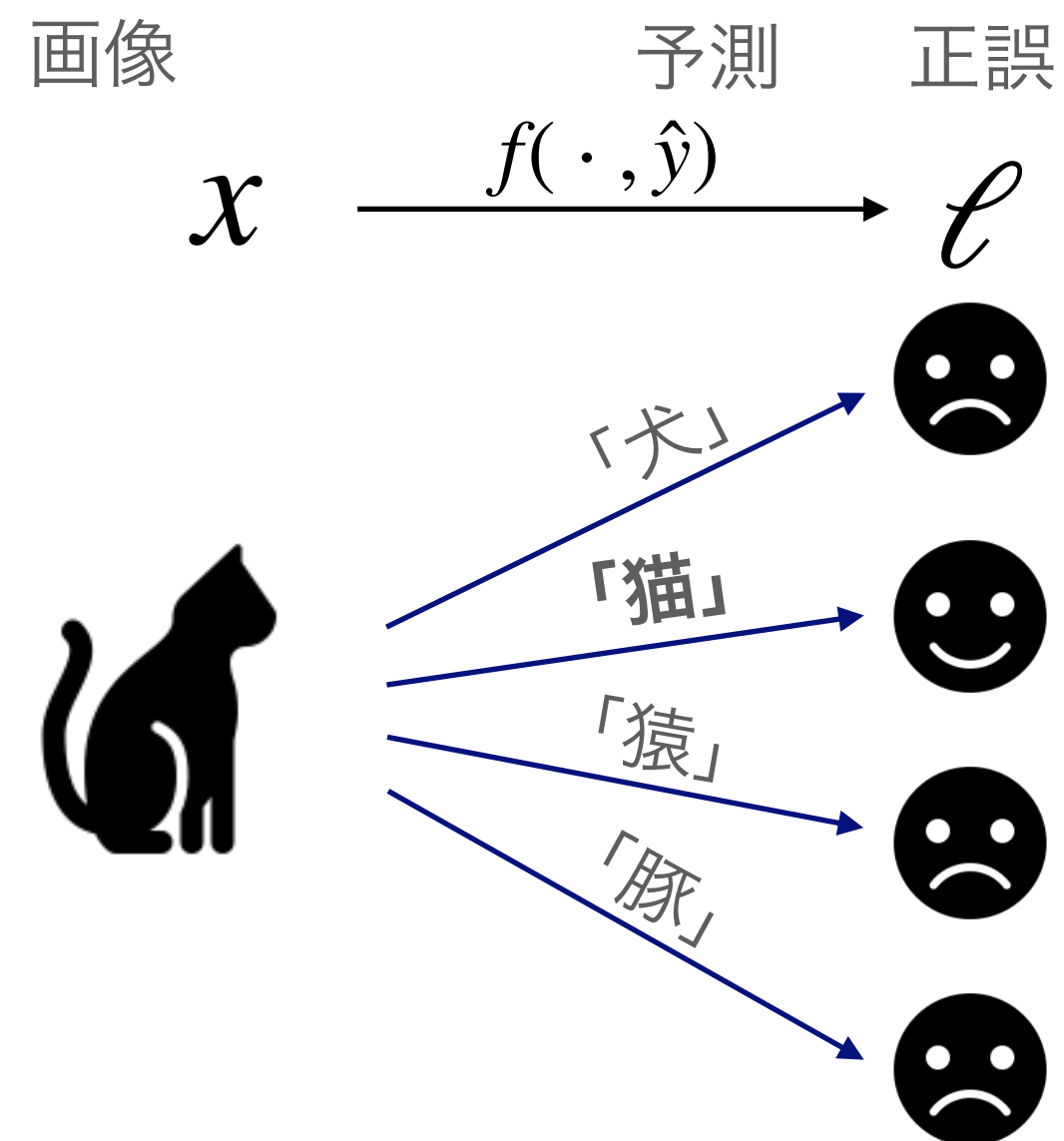
- データから意思決定に資する規則性を抽出するには？
 - 意思決定の仕方を変えても不変の規則性が必要
- 世界はモジュール的である（と仮定）
 - ある部分の意思決定の仕方を変えても
（例：年齢が若くても運動する）
他の部分の因果機序（ $y = f(x, a)$ ）は不変。介入は局所的で、
介入した変数以外の関係は変わらない（モジュール性）
- 世界のモジュール性に沿ったモデリングをする
＝因果推論



教師あり機械学習（普通の予測）との違い：例

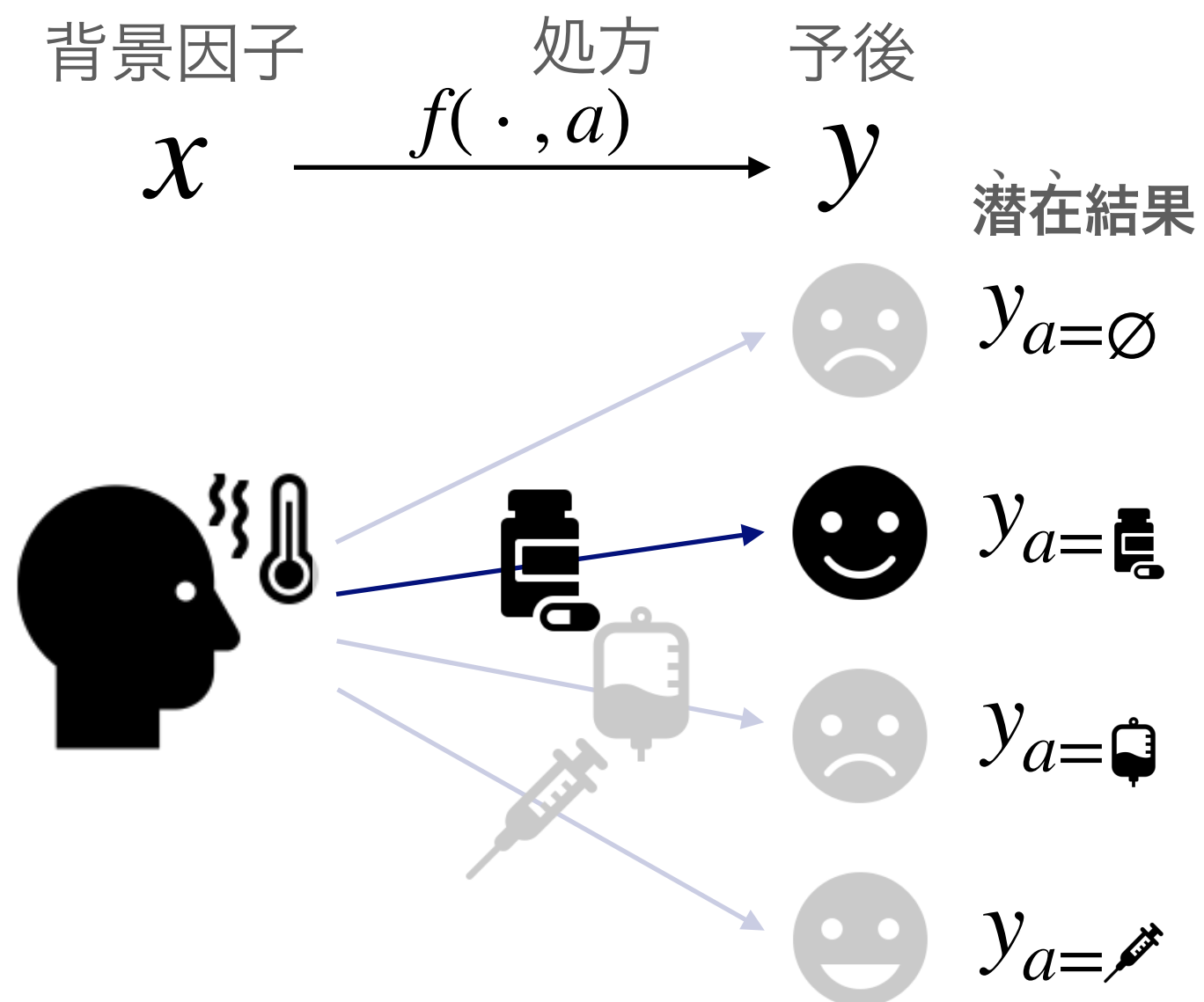
意思決定問題では一般に反事実的結果が欠損

教師あり学習



- 画像認識
 - ラベル以外は間違い

因果推論

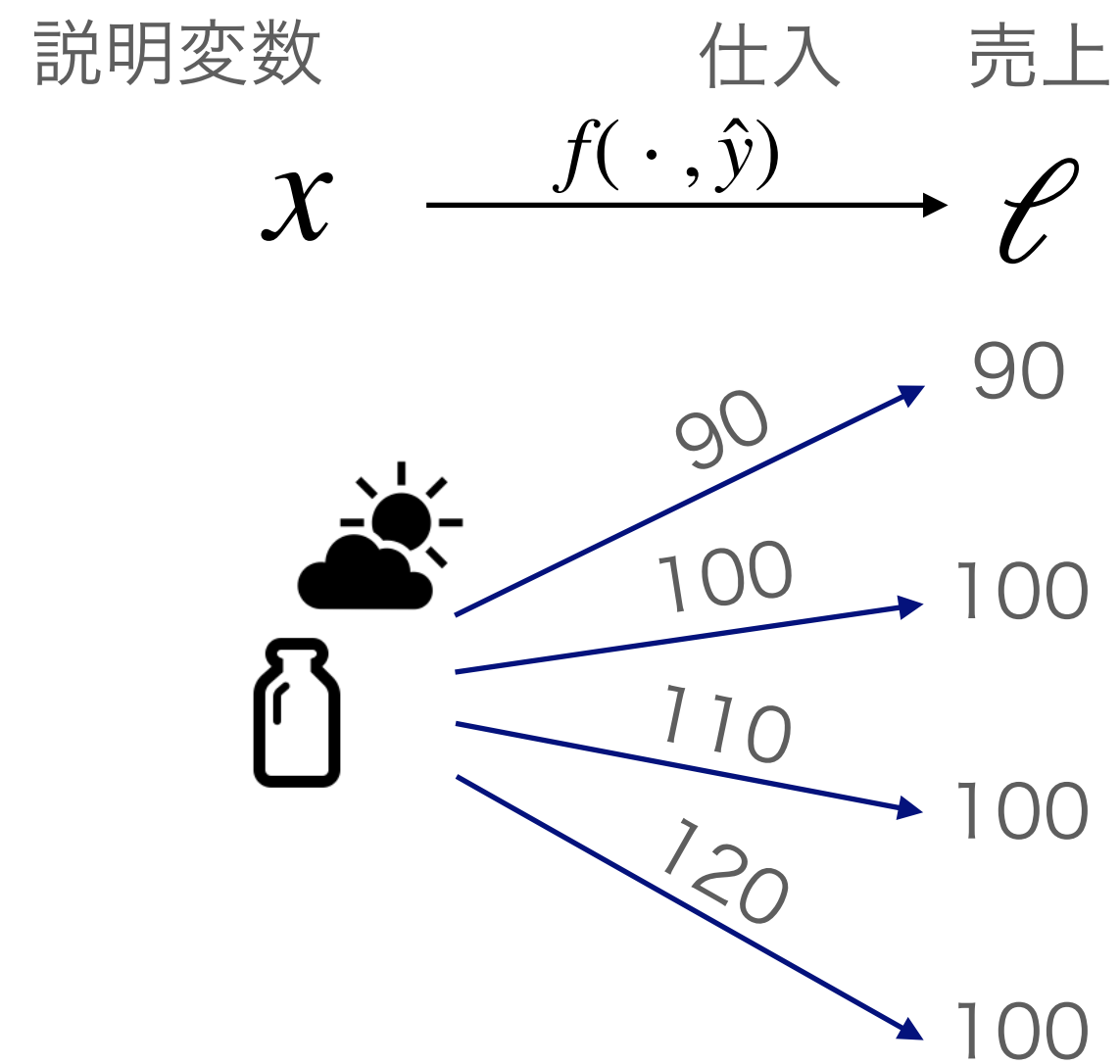


- 医療措置予後予測
 - 実際に行った措置の結果のみ観測

教師あり機械学習（普通の予測）との違い：実践的な例

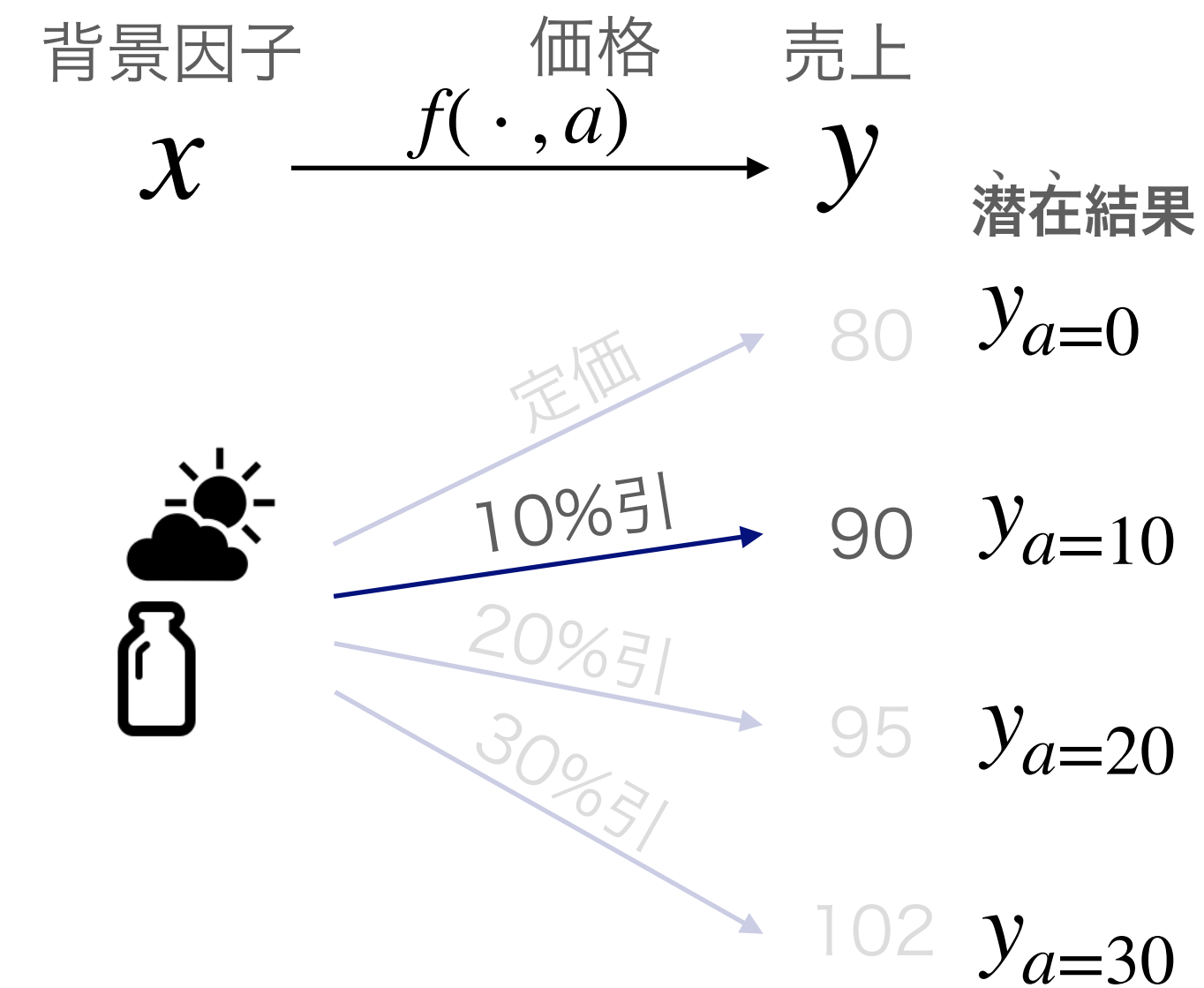
発注最適化＝(ほぼ)教師あり、価格最適化＝部分観測

教師あり学習



- 需要予測
 - 売れた量＝需要と仮定

因果推論

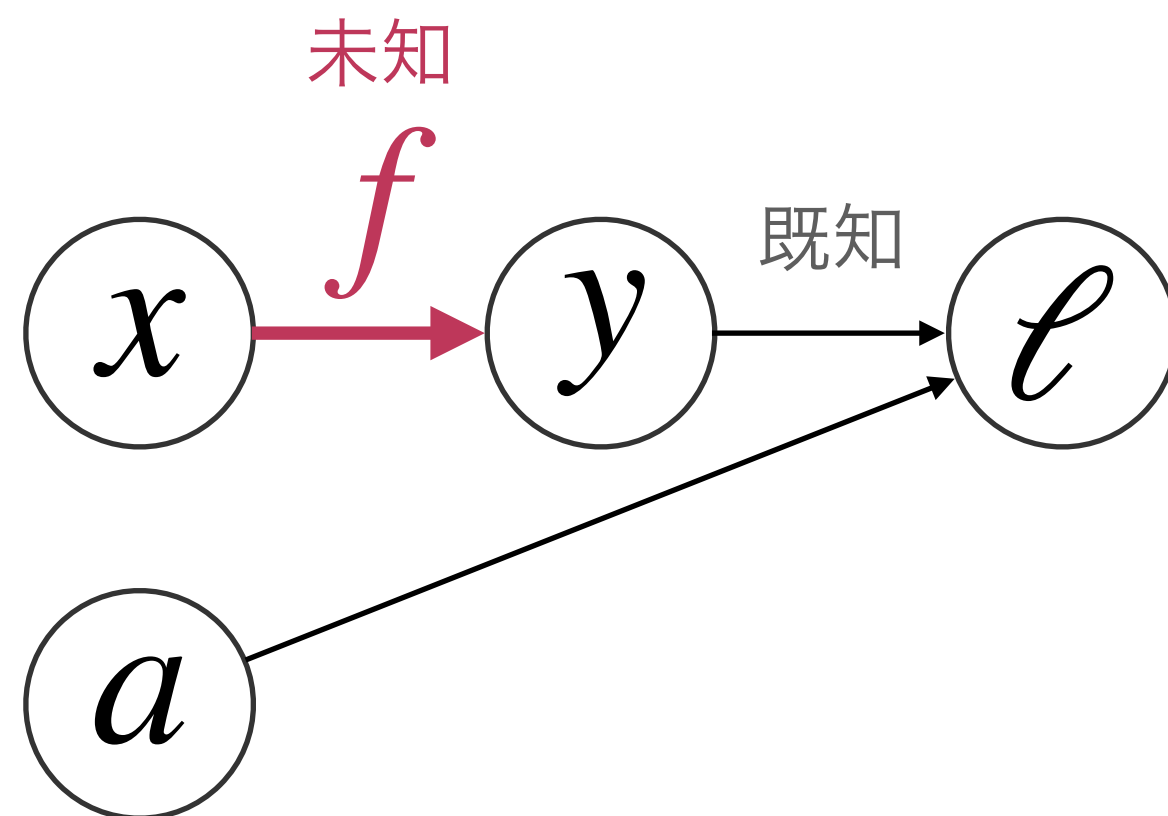


- 価格弾力性推定
 - 実売価格に対する需要のみ観測

教師あり機械学習（普通の予測）との違い：一般化

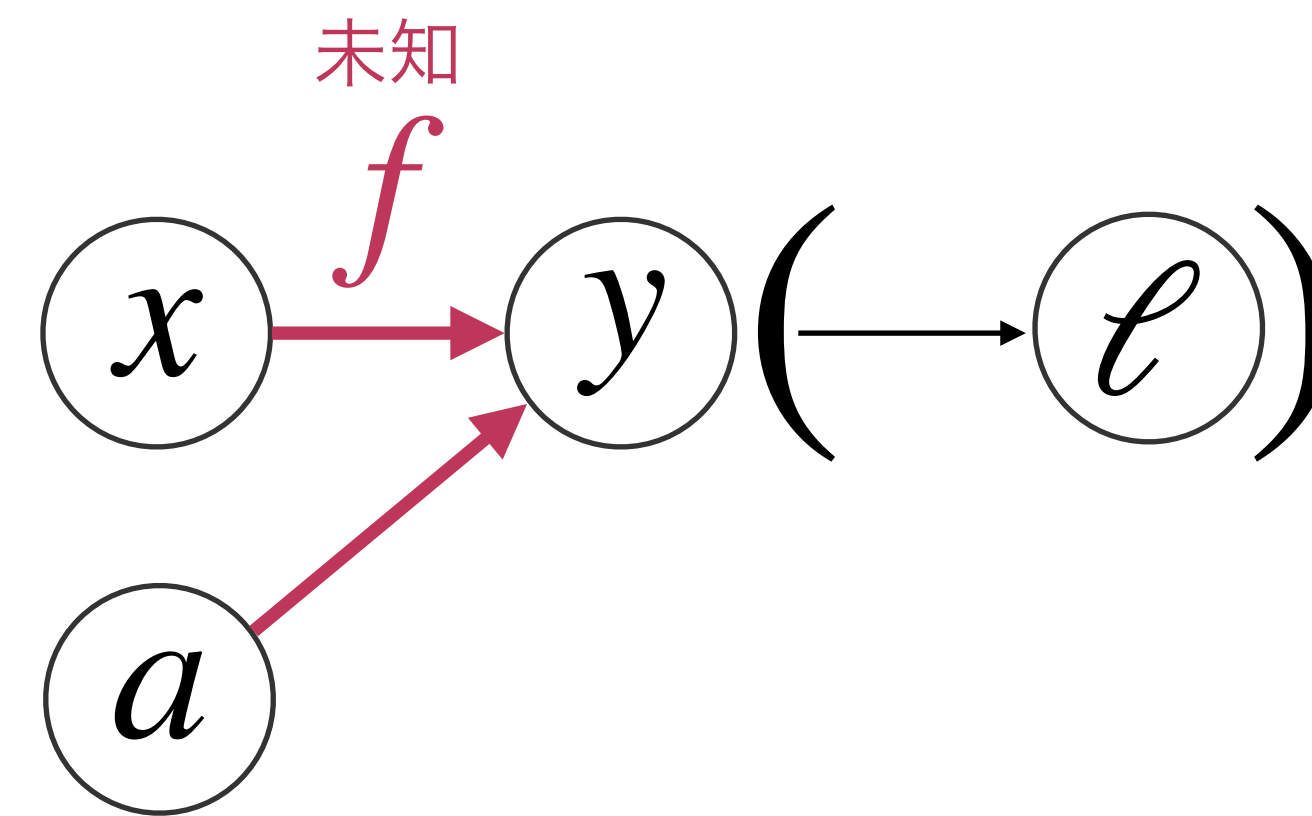
最終的な意思決定における変数が推定対象に直接関与するか

教師あり学習



- 意思決定変数 a と未知関数が分離

因果推論



- 意思決定変数 a を入力に含む関数が未知

(教師あり) 機械学習の概説

教師あり学習を応用して因果推論を行う

データから規則性を抽出し、予測や意思決定を行う技術

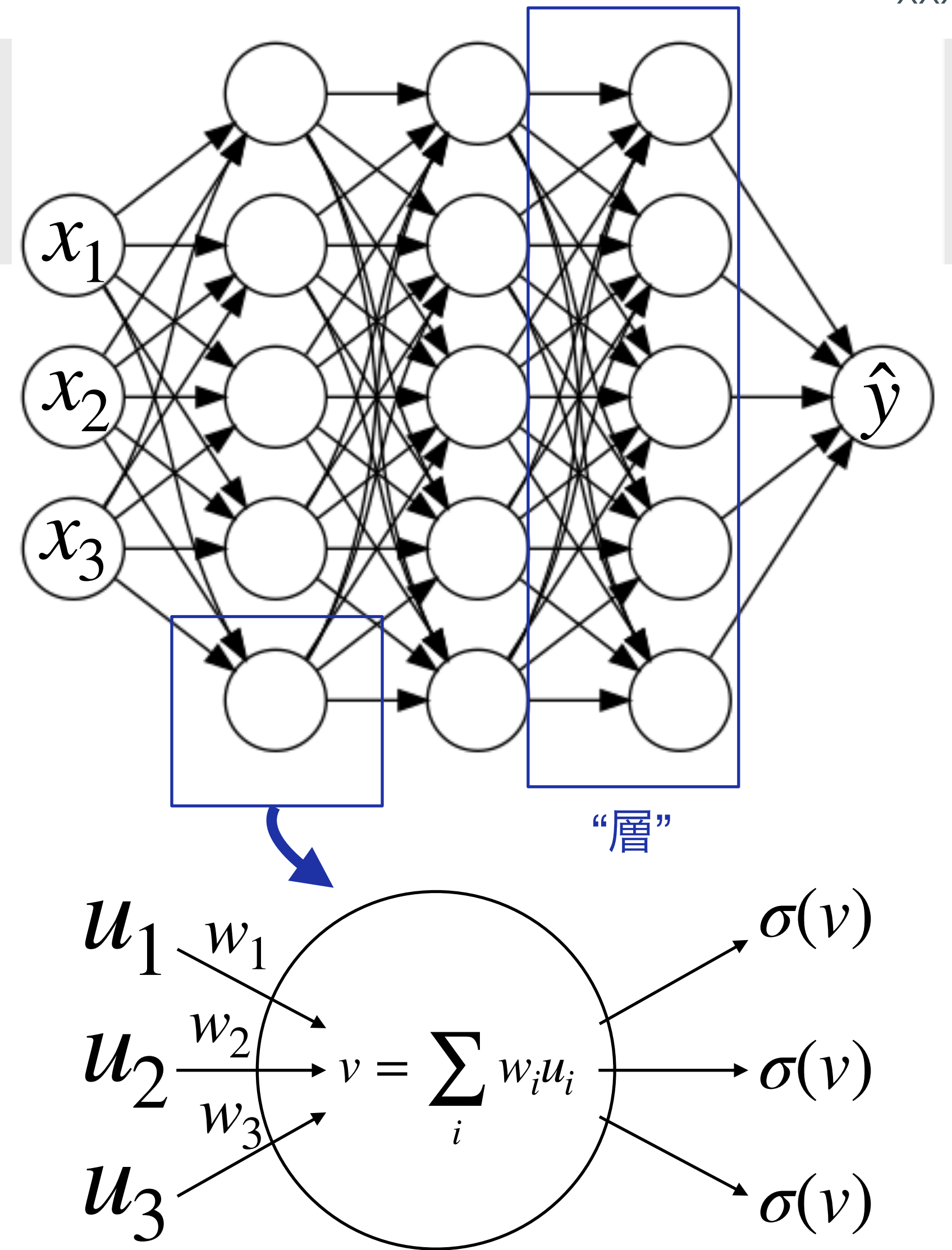
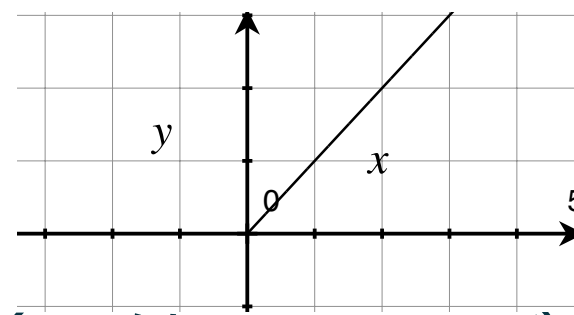
- 訓練データ D から予測モデル（関数） $\hat{f}$ を推定する学習器 A を考える
 - $\hat{f} = A(D)$
- 予測モデル $\hat{f}$ は変数 x をもとに変数 y を予測する： $y \doteq \hat{f}(x)$ と期待される
- 学習器 A は通常何らかの損失関数 L （予測誤差の近似）を最小化する関数として定義
 - $\hat{f} = \arg \min_{f \in \mathcal{F}} \hat{L}(f; D)$ ： $\mathcal{F}$ の中でデータ D に基づく損失関数 $\hat{L}$ を最小化する関数を $\hat{f}$ とする
 - $\mathcal{F}$ （花文字の F ）は仮説集合（モデルクラス）と呼ばれる。ハット $\wedge$ は推定量を表す。
 - 損失関数にハットを付けることもある。有限のデータで計算する経験損失（期待損失の推定量）だから
- 仮説集合 $\mathcal{F}$ ・損失関数 L ・最適化アルゴリズムを別々に考え、組み合わせられることが多い
 - 機械学習では、非線形で柔軟なモデルを考え、それによる個別化された意思決定を志向する傾向
 - ※古典的な統計学との重なりも大きい

具体例（仮説集合）

多層パーセプトロン

- 深層モデルの基本形
- ベクトルの線形変換と要素ごとの非線形変換 σ を繰り返し適用
 - $f(x) = W_L \sigma (\dots W_2 \sigma (W_1 x) \dots)$
- $\theta = (W_1, \dots, W_L)$ がモデルパラメータ（推定対象）
- 非線形関数は正規化線形ユニット（Rectified linear unit; ReLU）
 $\sigma(v) = \max(0, v)$ など
- 判別問題の場合は最終層にsoftmax関数（二値ならシグモイド関数）

$$g(u)_i = \frac{e^{u_i}}{\sum_j e^{u_j}} \text{ を適用}$$



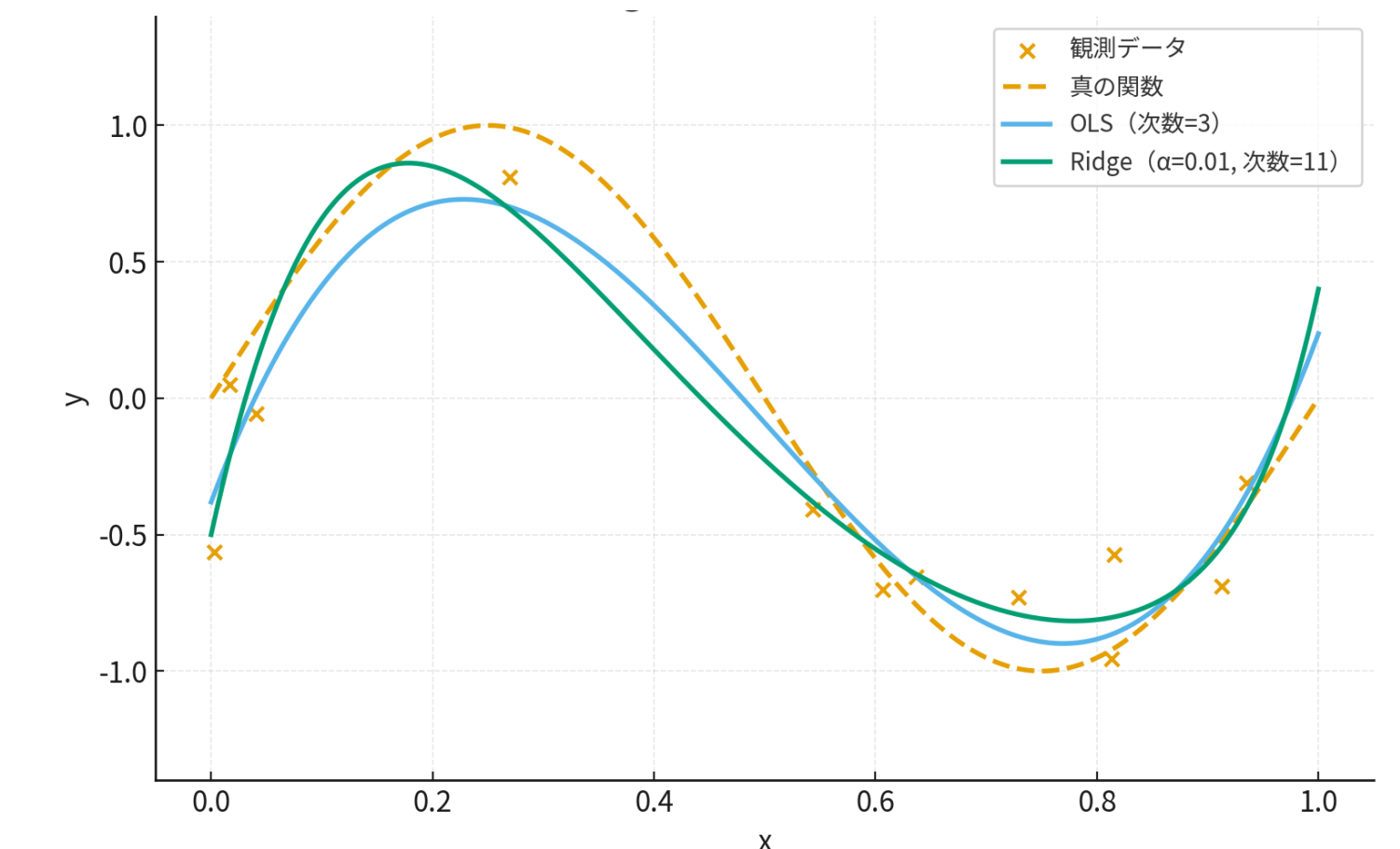
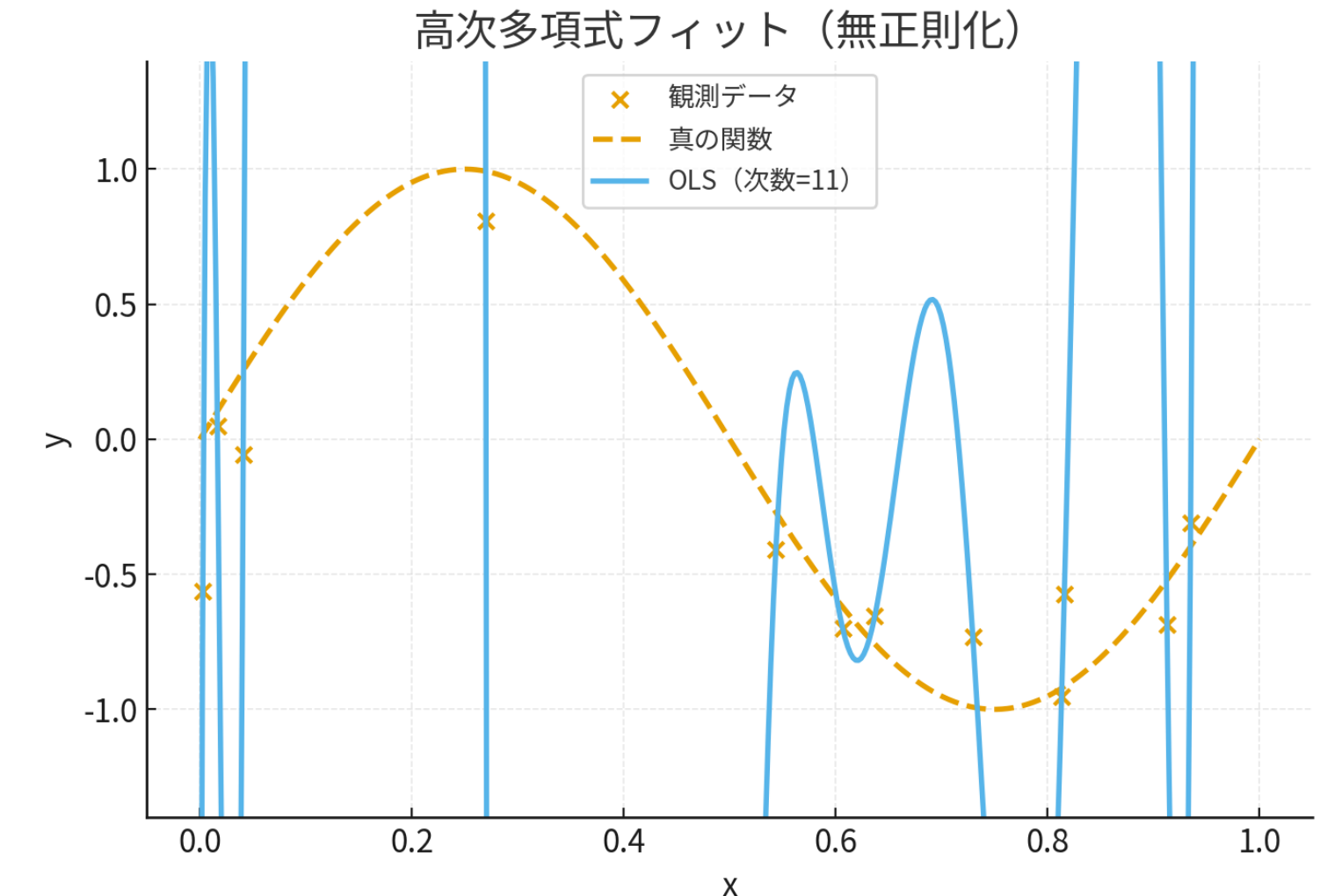
具体例（損失関数）

予測精度が高く、簡潔な関数を選ぶ

- 個々のデータ点に対する損失 ℓ の典型は
 - 回帰は2乗誤差： $\ell(f; o^i) = (y^i - f(x^i))^2$
 - 二値判別は交差エントロピー：
 $\ell(f; o^i) = -y^i \log f(x^i) - (1 - y^i) \log(1 - f(x^i))$
- モデルの“複雑さ”に対する罰則を加える場合が多い

$$\hat{L}_{\text{reg}}(f) = \frac{1}{N} \sum_{i=1}^N \ell(f; o^i) + \alpha R(f)$$

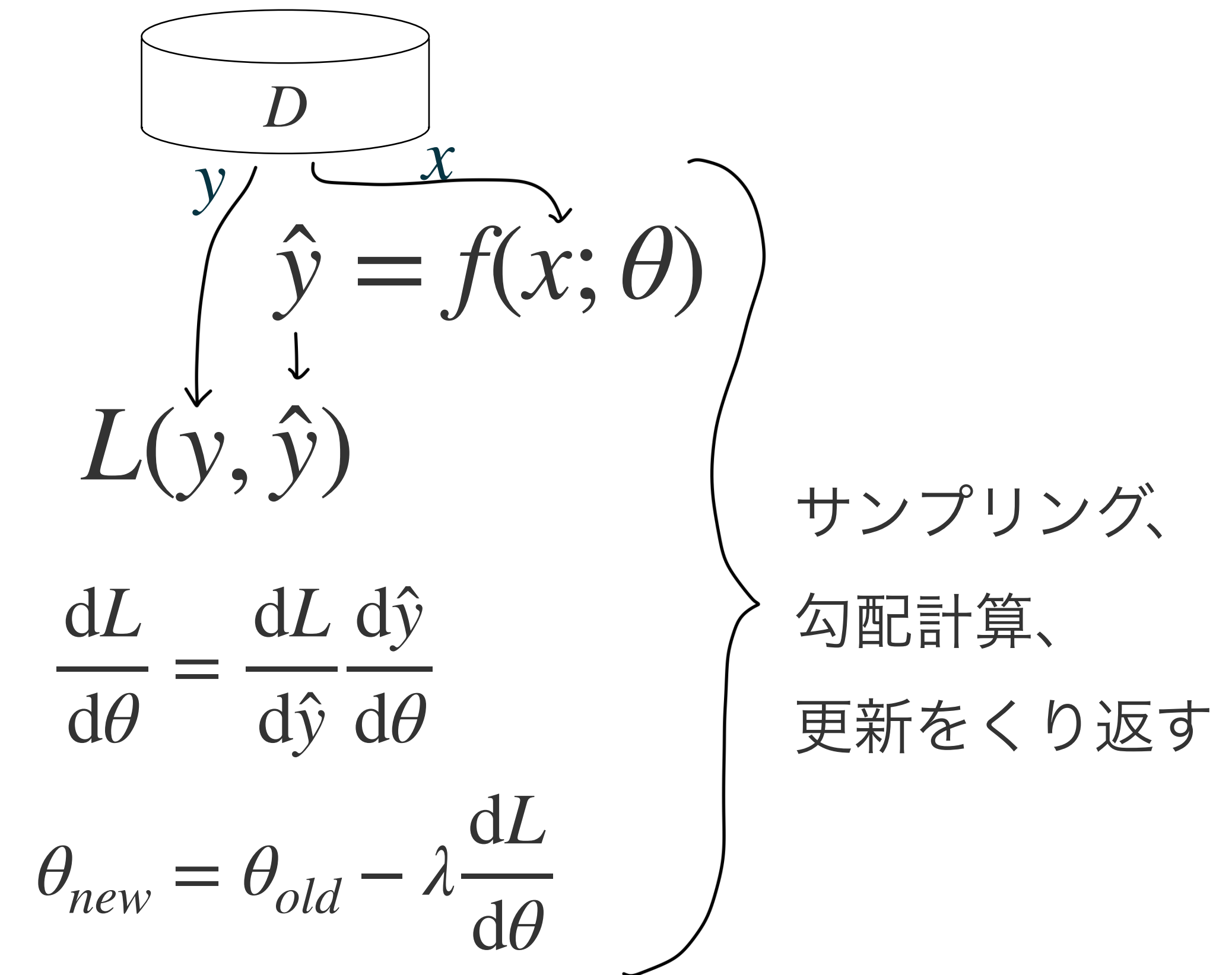
- 罰則はパラメータのノルムの2乗など $R(f) = \|\theta\|_2^2$ （リッジ正則化と呼ばれる）
- サンプルサイズに対して仮説集合が複雑すぎると過学習するため
 （訓練データに過剰に適合し、未知のデータに対する精度が下がる）。罰則の強さは検証用データで選ぶ（モデル選択）



具体例（最適化アルゴリズム）

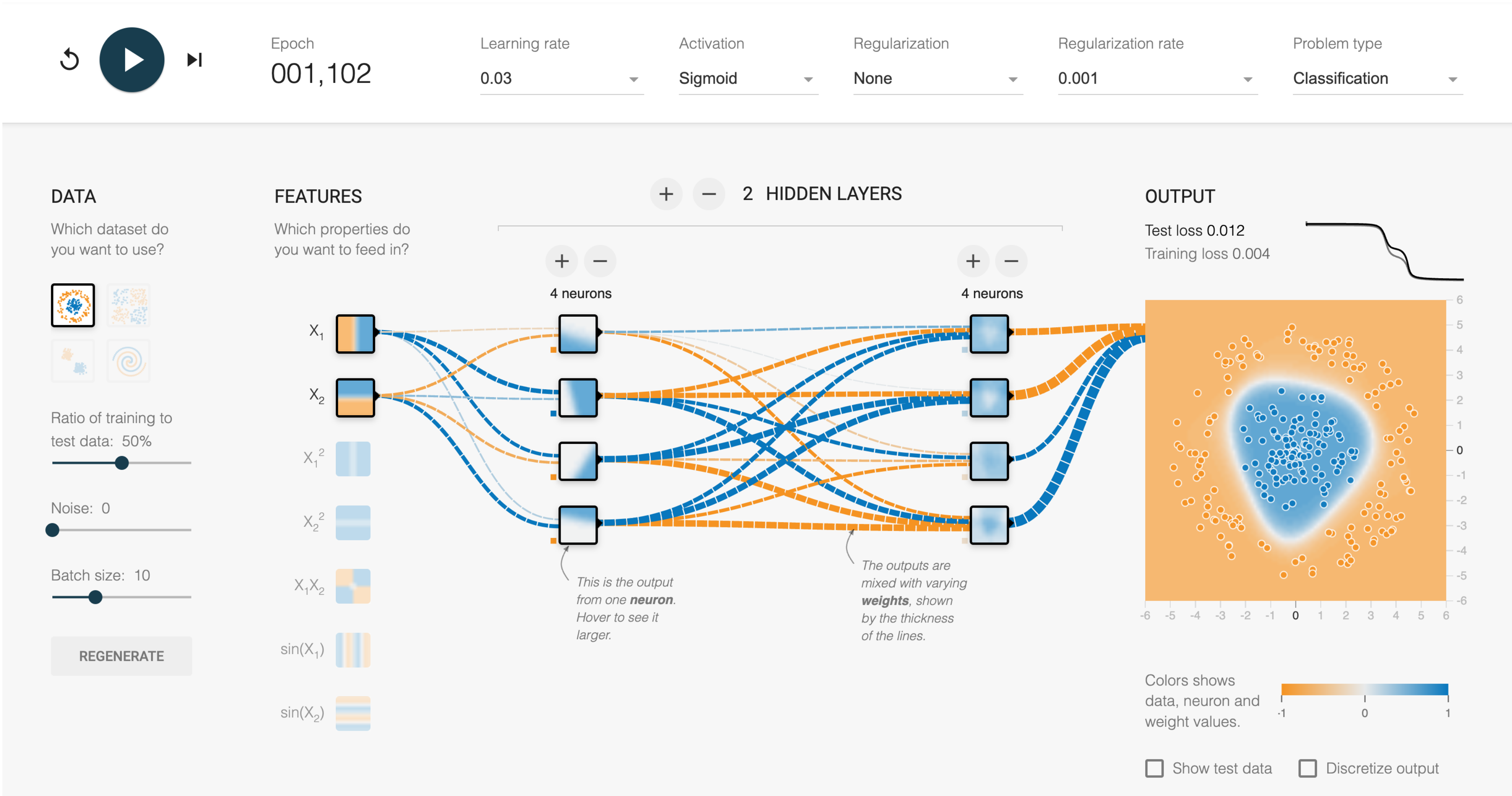
（確率的）勾配降下法

1. 仮説集合 $\mathcal{F}$ 、損失関数 L 、データ D を入力
 - D は入力変数 x と予測対象 y を含む
2. データの一部をランダムに取得
3. L を最小化するように勾配の逆方向に
少しだけ θ を更新
4. 上記2、3を繰り返す
 - 2の抽出を非復元で行って全データを学習した
1回分をepochと呼び、epochをさらに繰り返す
 - 訓練データとは別に取っておいた検証用データ
での精度が悪化し始めたら終了
5. パラメータ θ を出力



※ ベクトルでの微分は要素ごとの微分
のベクトル

多層パーセプトロンの学習例



A Neural Network Playground

おそらくだいたい正しい (PAC) ことを保証する

- 0-1 損失を用いた二値判別を $d - 1$ 次元線形モデル (定数項を含めてパラメータ数は d) で行うとき、
- 仮説集合 $\mathcal{F}$ の中で最善のモデルに比べて高々 ε 程度しか悪くない
- ような結果をもたらすデータが出る確率が $1 - \delta$ 以上、つまり
- $$p \left(\mathbb{E}[\ell(\hat{f}; o)] - \min_{f \in \mathcal{F}} \mathbb{E}[\ell(f; o)] \leq \varepsilon \right) \geq 1 - \delta$$
- となるのに必要なサンプルサイズ $N_{\mathcal{F}}(\varepsilon, \delta)$ は、定数 C_1 、 C_2 が存在して、以下が成立
- $$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq N_{\mathcal{F}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}$$