化学データに対する
データ解析入門

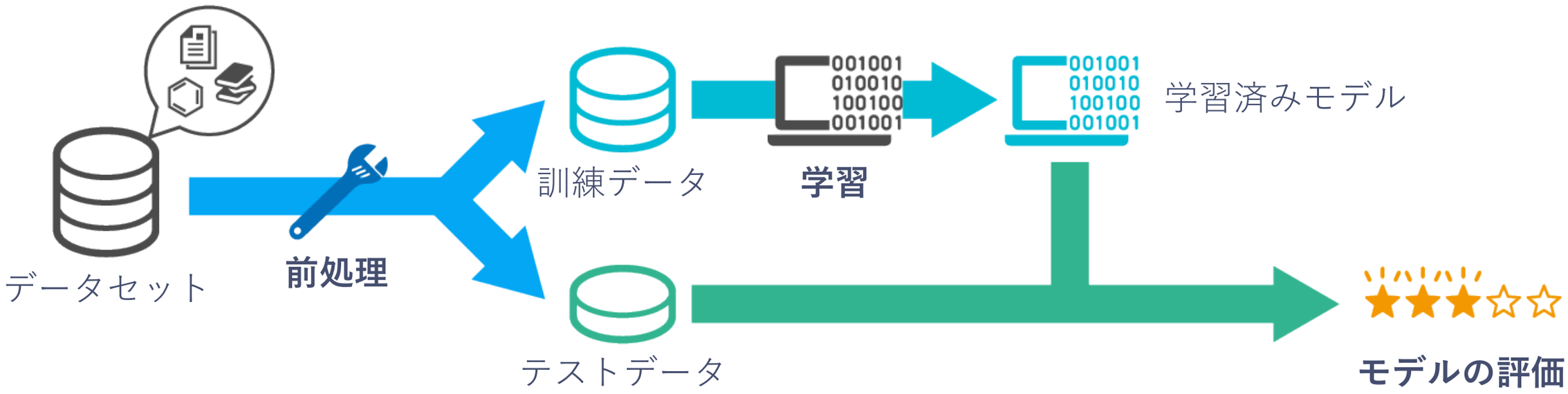
井上 貴央, 加藤 涼太, 沼田 康平

2019/10/28 @ ケモインフォマティクス若手の会

やること

以下の内容をやります:

- **Section 0**: 機械学習とは?
- **Section 1**: データの前処理
- **Section 2**: モデルとその学習
- **Section 3**: モデルの評価

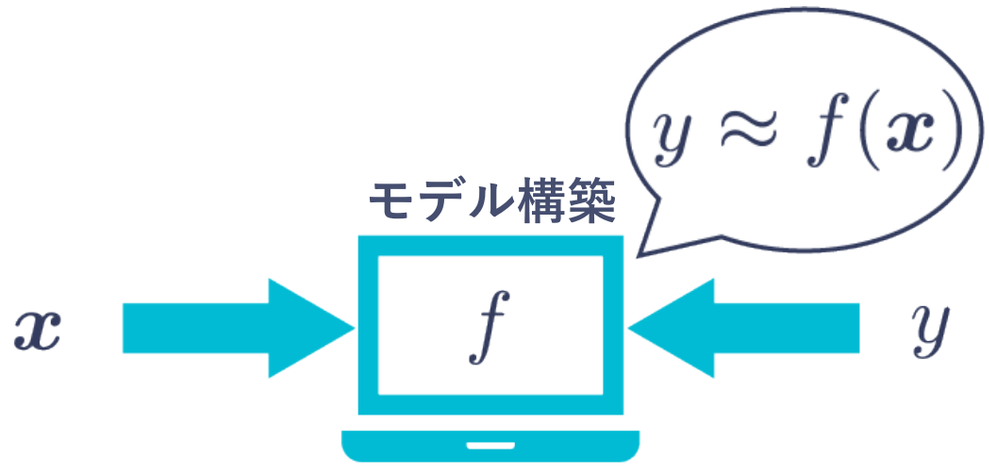


機械学習とは?

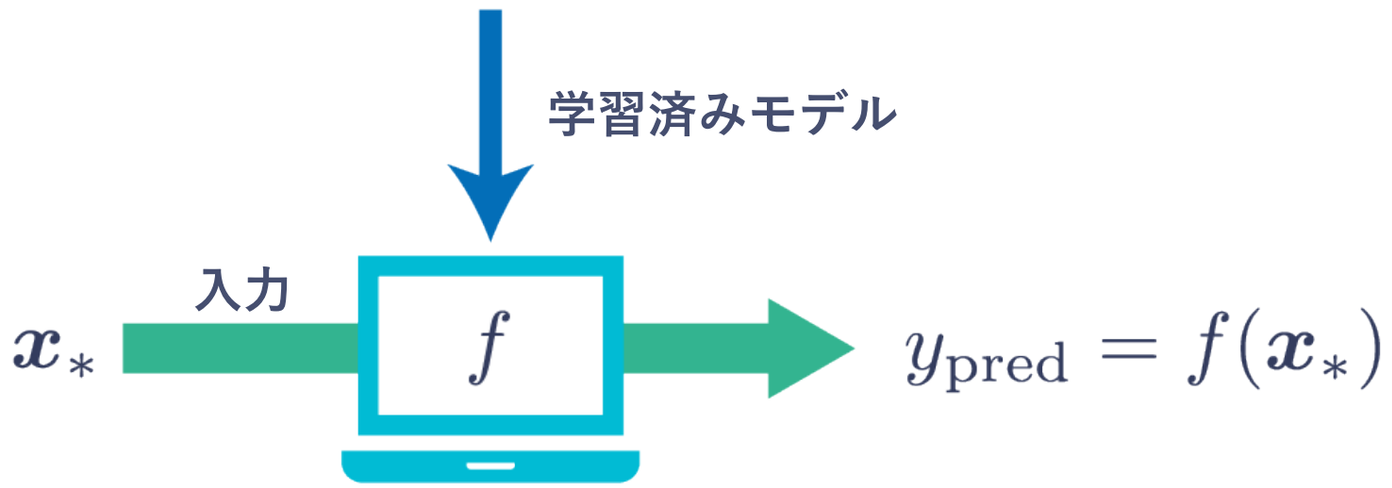
既存のデータをうまく説明できる**モデル** (関数 f) を学習し,

未知データに対する予測などを可能にすること

学習 (training)

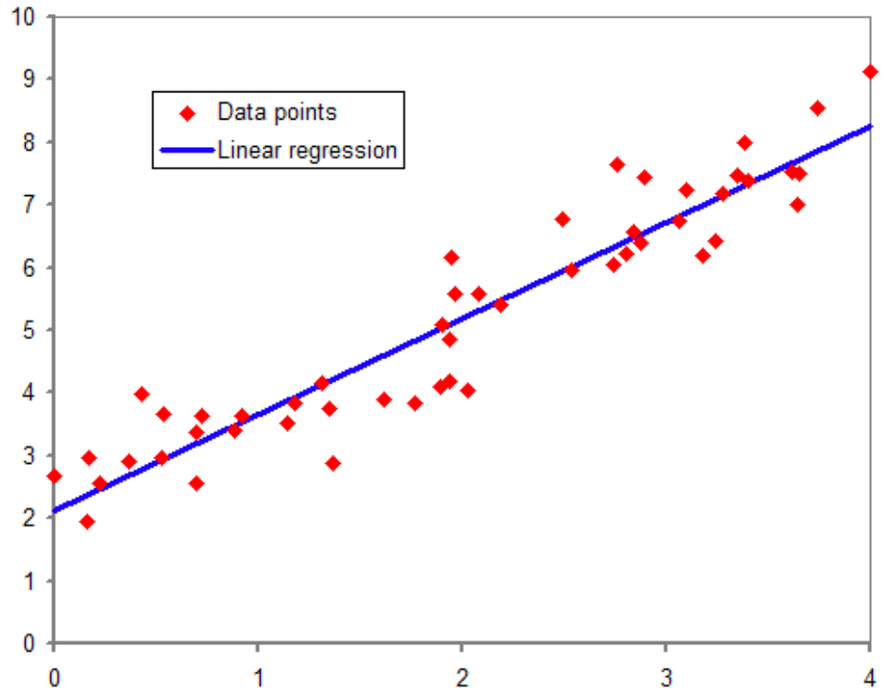


予測 (prediction)



機械学習の主なタスク

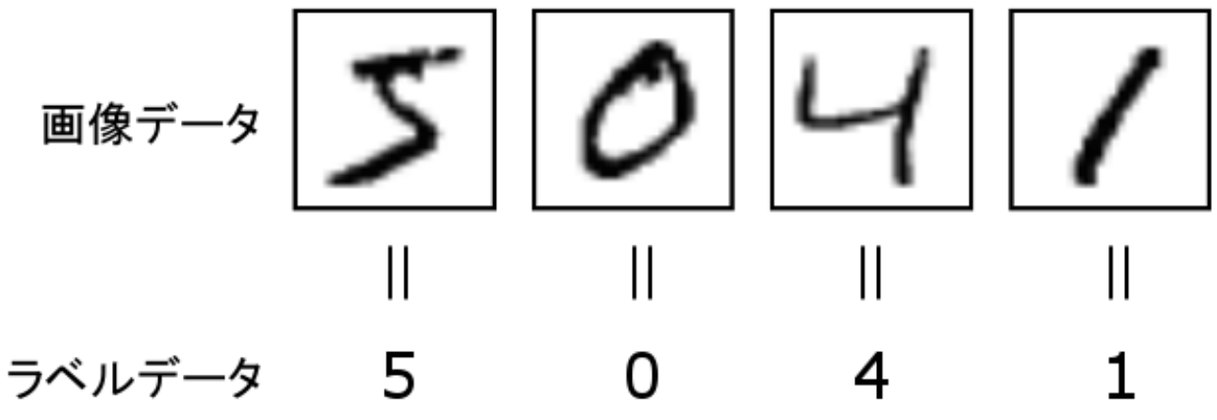
- **回帰** (regression) ← **今日やるのはこちら**



← こんなの

(Wikipedia: 線形回帰の項の図を引用)

- **分類** (classification)



← 入力

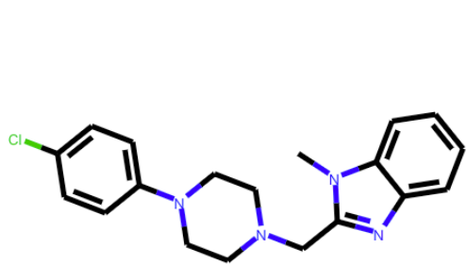
← これを当てる

https://weblabo.oscasierra.net/python/ai-mnist-data-detail.html

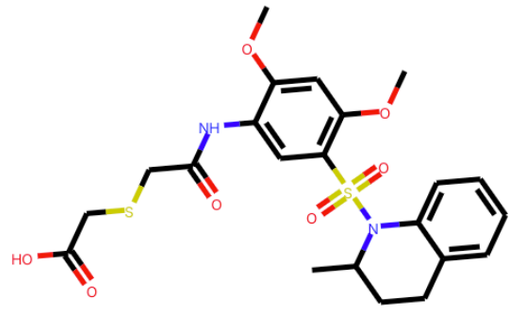
今回利用するデータ

Lipophilicity Data (from MoleculeNet[1])

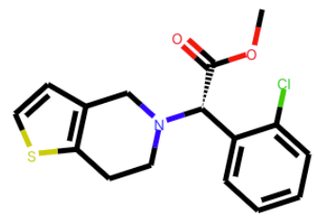
- オクタノール-水分配係数 (pH=7.4 での logD) の実験値
- データ数: 4200



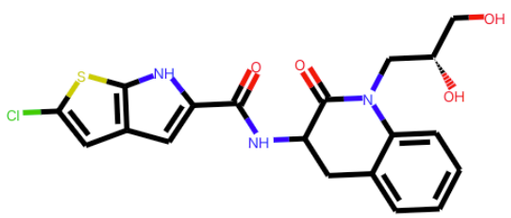
logD = 3.54



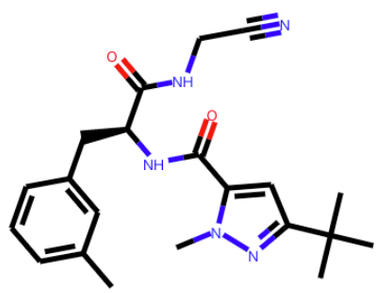
logD = -1.18



logD = 3.69



logD = 3.37



logD = 3.1

[1] MoleculeNet (available at: http://moleculenet.ai)

Part 1

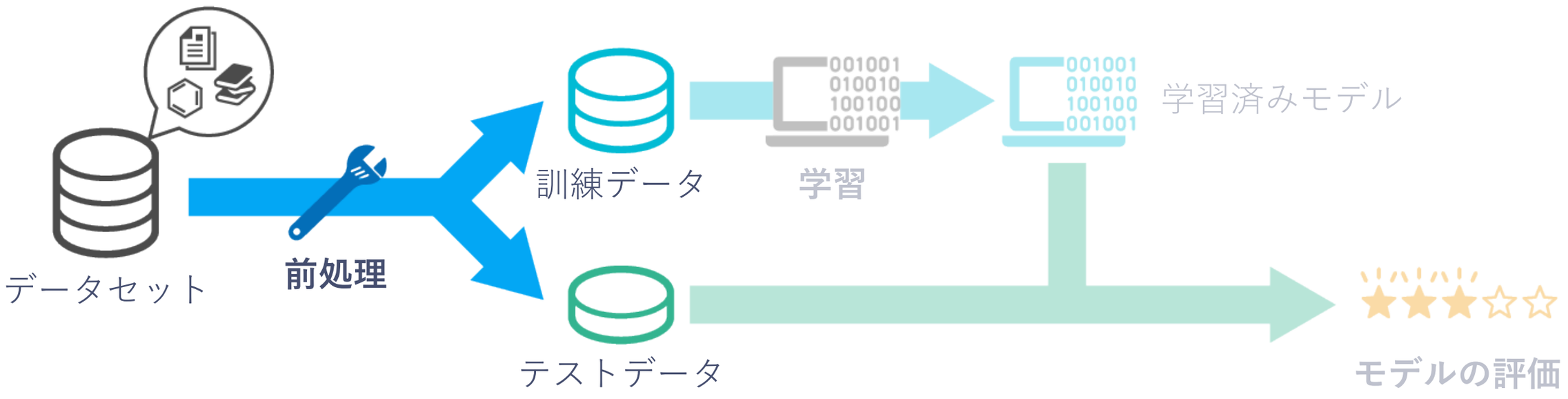
データの前処理

前処理とは?

前処理 (preprocessing): データを解析する前にデータを整形すること

前処理の目的

- モデルへの入力を数値化するため
- 欠損値を処理するため
- モデルの精度を向上させるため
- モデルの評価をするため



入力を数値化する

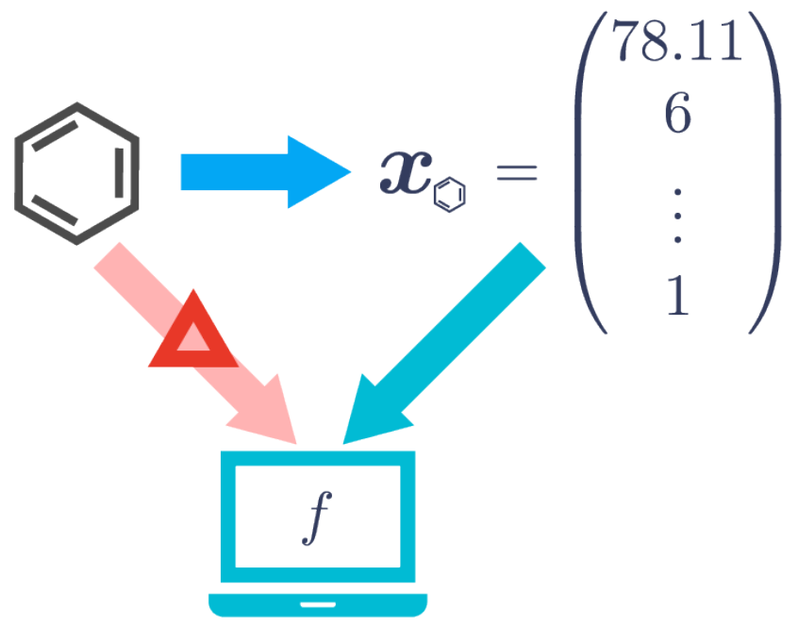
分子構造を直接モデルに入力するのは難しい.....

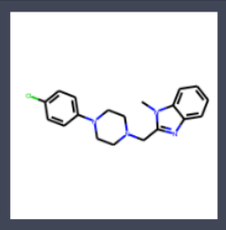
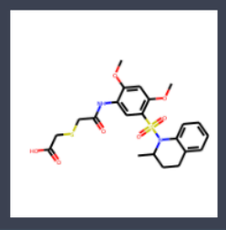
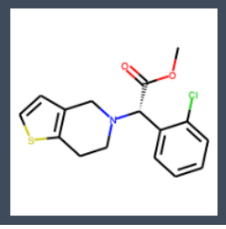
➔ 分子構造を数値データに変換しよう

記述子 (descriptor): 分子構造から算出される数値

ex) 分子量, 炭素原子の数, ベンゼン環の数, ...

例えば, RDKit[1]を利用して計算できる



	ROMol	MolWt	ExactMolWt	HeavyAtomCount	NumAliphaticRings	NumAromaticRings	MinPartialCharge	MaxPartialCharge	fr_ben
0		340.858	340.145474	24	1	3	-0.368964	0.123343	2
1		494.591	494.118143	33	1	2	-0.495171	0.312967	2
2		321.829	321.059027	21	1	2	-0.467586	0.327301	1

[1] http://www.rdkit.org

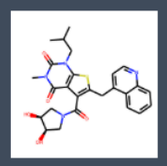
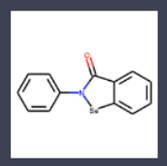
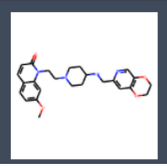
欠損値を処理する

記述子が計算できないものもデータには含まれているかもしれない……

➔ 欠損値を何らかの方法で除去しよう

処理の例:

- サンプルを除去する
- 記述子を除去する
- 平均値/中央値/最頻値/0 で補完する

	ROMol	MolWt	ExactMolWt	HeavyAtomCount	NumAliphaticRings	NumAromaticRings	MinPartialCharge	MaxPartialCharge	fr_benz
1560		508.600	508.178041	36	1	4	-0.388500	0.331336	1
1561		274.181	274.984935	16	0	3	NaN	NaN	2
1562		450.539	450.226705	33	2	3	-0.496687	0.250557	1

➔ 計算できない記述子があった1サンプルを除去した

精度を向上させる

モデルの精度を低下させうる記述子があるかもしれない……

➔ 利用する記述子を選択しよう

要らない記述子の例:

- 記述子の値が一定

  ex) データセット内の分子に存在しないフラグメント記述子
- ある記述子と相関の高い記述子

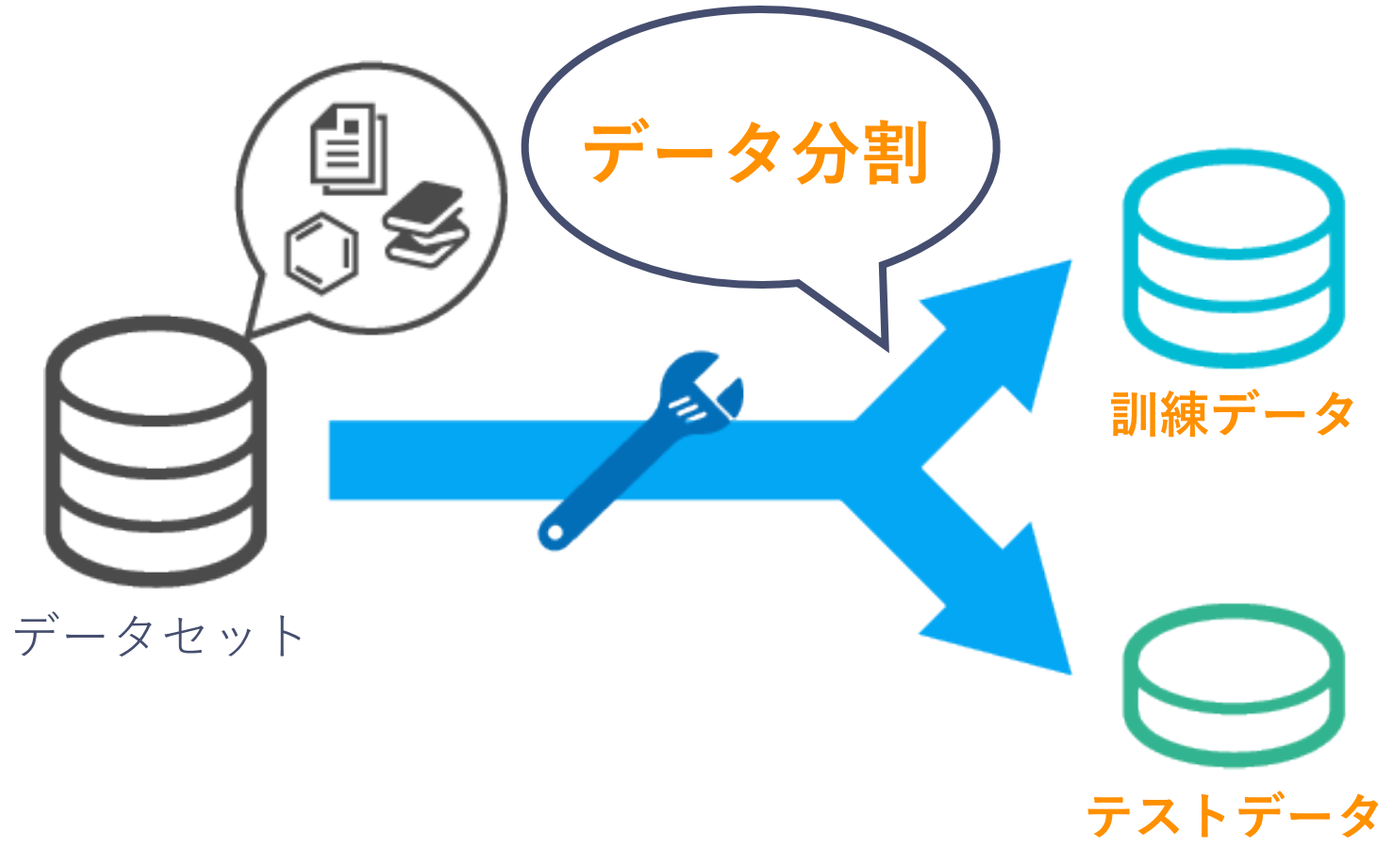
  ex) MolWt (平均分子量) と ExactMolWt (分子量)

➔ 値が一定の17記述子と, 別の記述子と相関の高い11記述子を除去した

モデルの評価をする

全データでモデルを学習させてしまうと, モデルの良し悪しを
評価できない.....

➔ 学習後のモデルを評価するためのデータを事前に取っておこう



➔ 訓練データ3359件とテストデータ840件に分割した

前処理完了!

Part 2

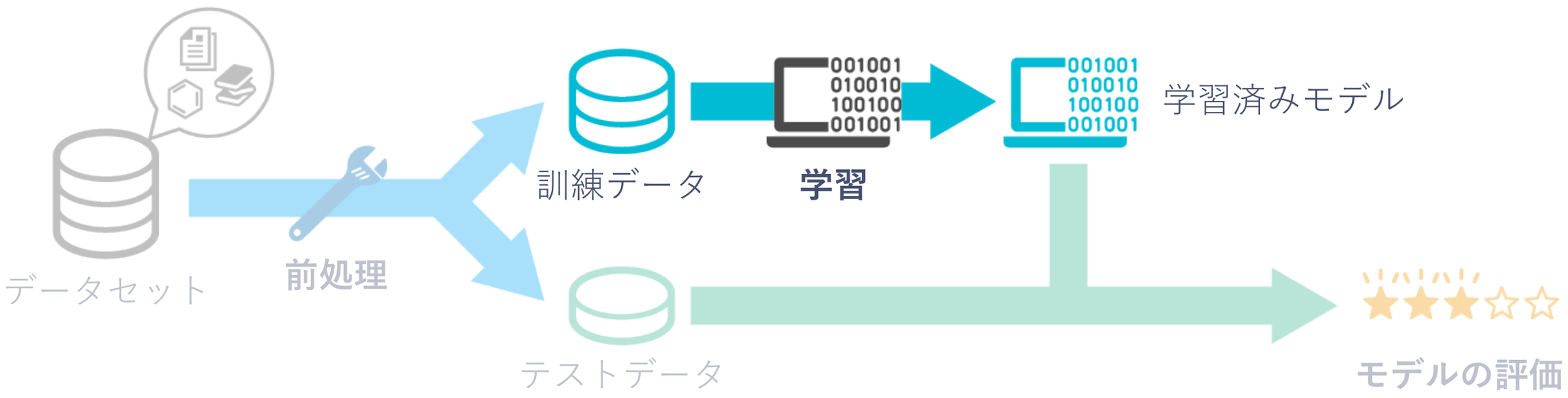
モデルとその学習

利用するモデルを決める

利用するモデルによって, うまく分子と物性値の関係性を
表現できるかどうかが変わる

今回検討するモデル:

- LASSO
- サポートベクター回帰 (Support Vector Regression: SVR)
- ランダムフォレスト (Random Forest: RF)



LASSO

- 線形回帰モデル:

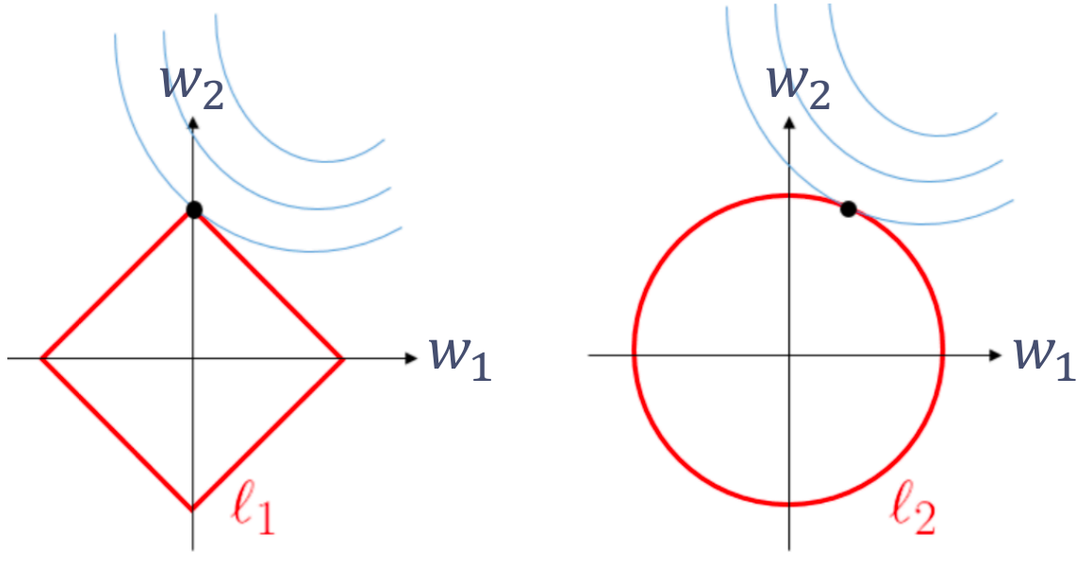
$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$$

- パラメータ $\boldsymbol{w}$ を以下の損失を最小化することで求める:

$$E(\boldsymbol{w}) = \underbrace{\sum_n (f(\boldsymbol{x}_n) - y_n)^2}_{\text{二乗誤差}} + \underbrace{\alpha \sum_i |w_i|}_{\text{正則化項}} \quad (\alpha\text{: 定数})$$

正則化項: パラメータ $\boldsymbol{w}$ が大きくならないようにし, 汎化性能を上げる
(未知データに対する予測性能)

- パラメータ $\boldsymbol{w}$ の成分が 0 になりやすい
(スパースなモデル)
- 定数 α は事前に設定する必要がある
(**ハイパーパラメータ**)
- データを平均 0, 分散 1 になるよう変換
(**標準化**) する必要がある



https://weblabo.oscasierra.net/python/ai-mnist-data-detail.html
図を一部改変

サポートベクター回帰 (SVR)

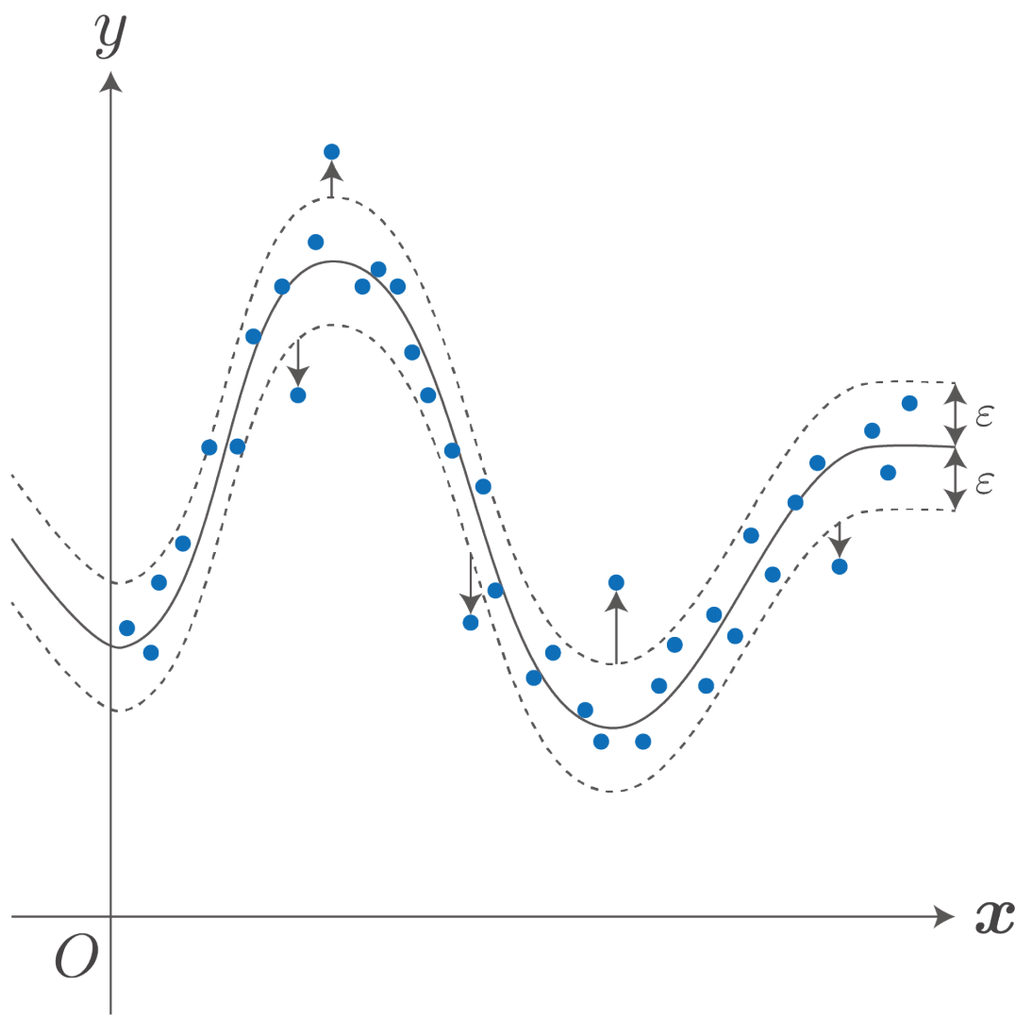
- 滑らかな非線形回帰モデル:

$$f(\boldsymbol{x}) = \sum_n \alpha_n \underbrace{\exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}_n\|^2)}_{\text{RBFカーネル}} + b$$

(γ: 定数)

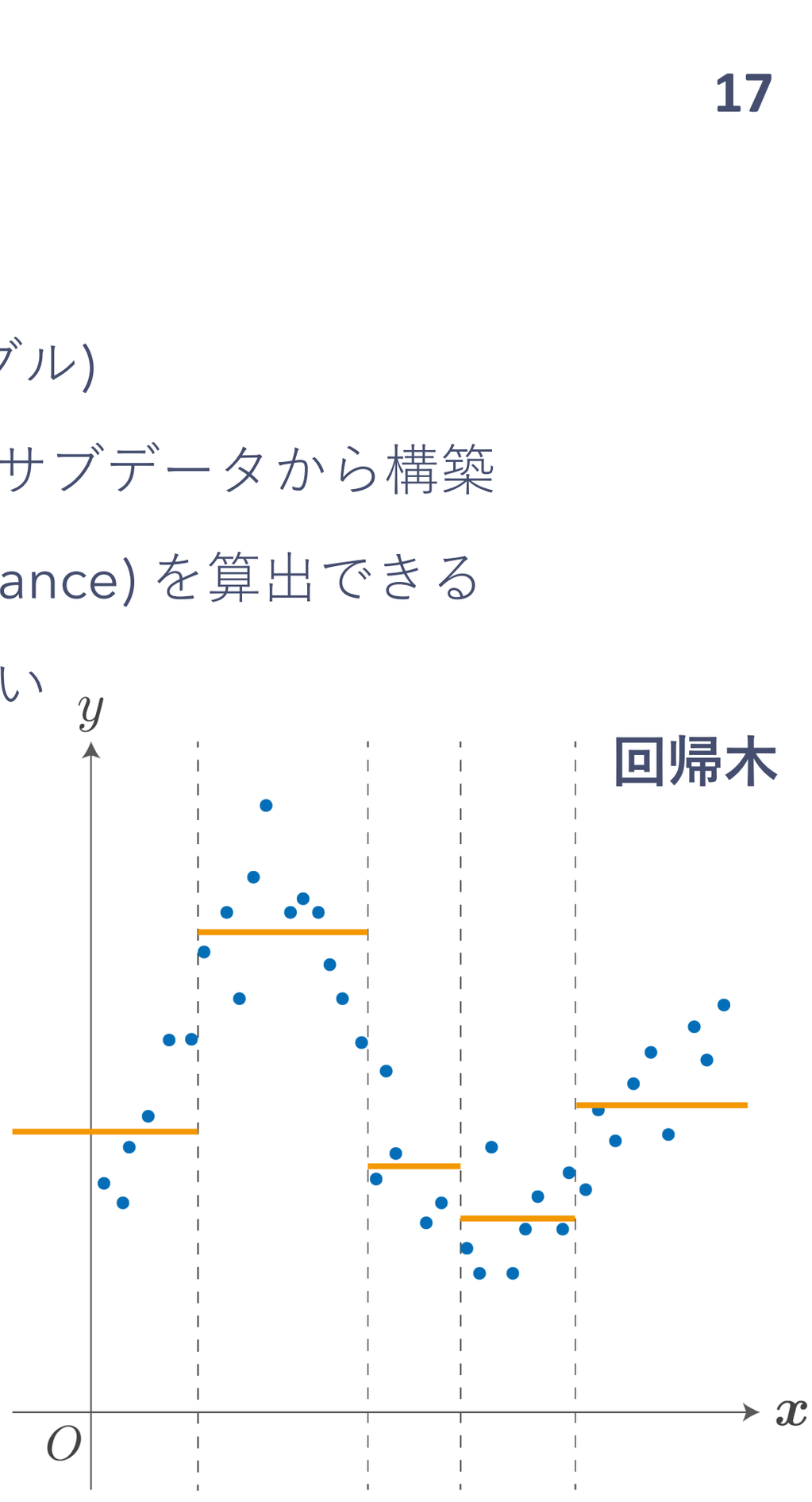
(α_n, b: 学習で決定)

- 回帰曲線から外れているサンプルでも, 誤差 ε 以内なら許容する
- 誤差が ε より大きいサンプルに対しては, 定数 C に比例したペナルティを課す
- ハイパーパラメータは三つ:
  - RBFカーネルの幅 γ
  - 許容誤差 ε
  - ペナルティ強度 C
- データの標準化が必要



ランダムフォレスト (RF)

- 不連続な非線形回帰モデル
- 複数の**回帰木**の出力値の平均 (アンサンブル)
  - 各回帰木をランダムサンプリングしたサブデータから構築
- 使用する記述子の**重要度** (feature importance) を算出できる
- ハイパーパラメータの調節がほぼ要らない
  - 主に回帰木の数を調節
- データの標準化が不要

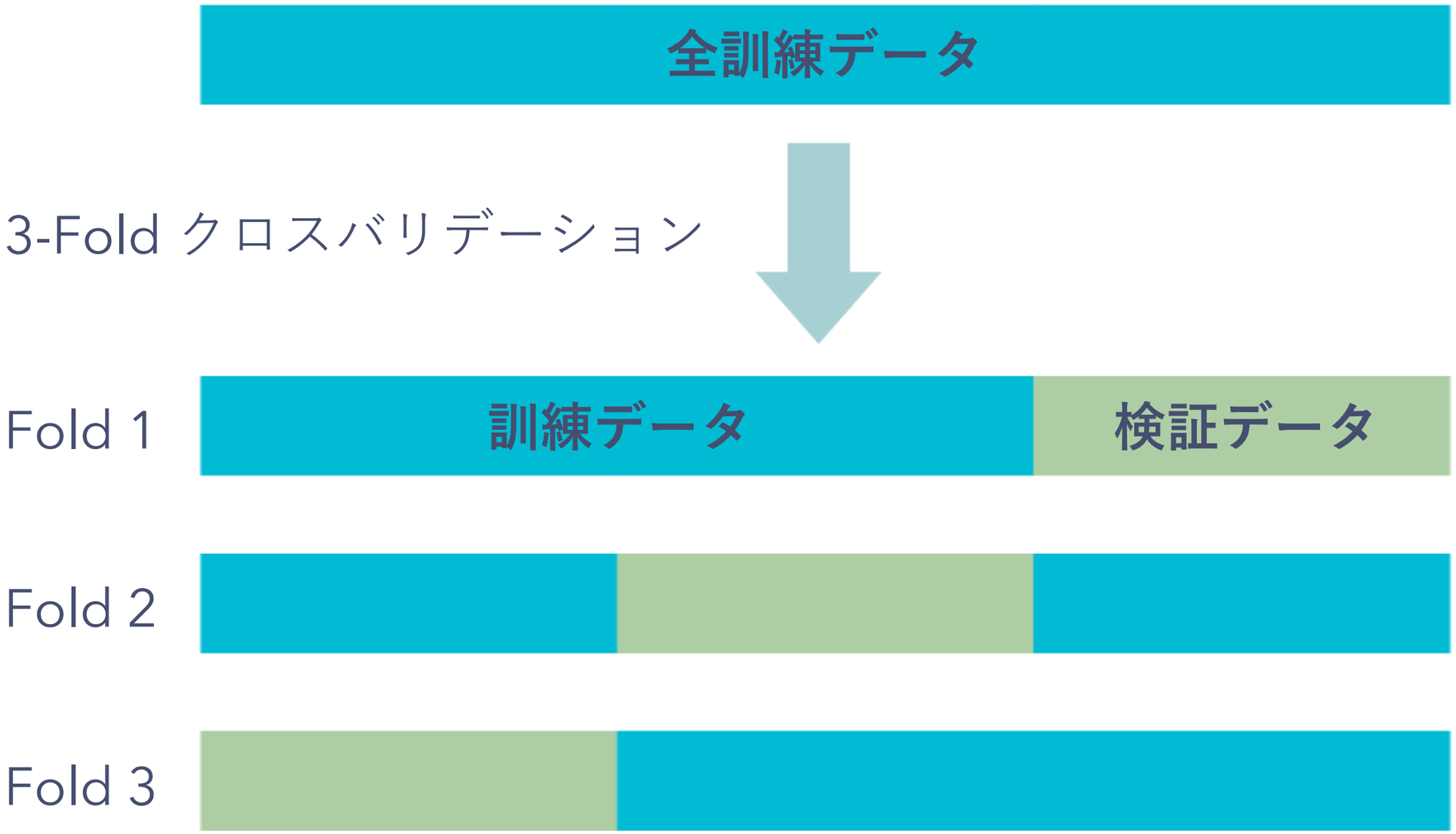


ハイパーパラメータの決定

モデルを確定させるには, ハイパーパラメータを決定する必要がある

➔ 気になるパラメータを全部試す (**グリッドサーチ**)

その際に**クロスバリデーション**を利用する



Now learning...

Part 3

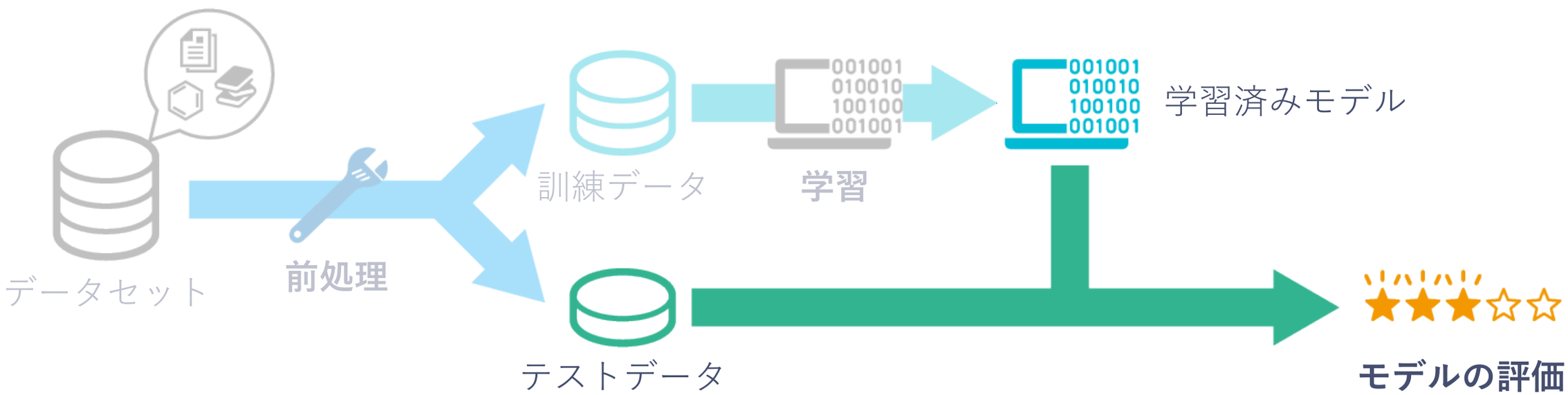
モデルの評価

どう評価するか?

とっておいたテストデータに対してモデルを適用し,
各サンプルに対する予測値と実測値のズレを確認する

確認する評価指標の例:

- 決定係数
- 平均二乗誤差の平方根
- 平均絶対値誤差



評価指標: 決定係数

n 番目のサンプル $(n = 1, \ldots, N)$ に対する予測値を $\hat{y}_n$, 実測値を y_n,
実測値 y_n の平均値を $\bar{y}$ とする

決定係数 (coefficient of determination):

$$R^2 = 1 - \frac{\sum_n (y_n - \hat{y}_n)^2}{\sum_n (y_n - \bar{y})^2}$$

- モデル同士を比較する際に利用する指標
- モデルがデータをどれだけ表現できるかの度合いを表す
- 1以下の値をとり, 1に近いほどモデルの当てはまりがよい

評価指標: 誤差

n 番目のサンプル $(n = 1, \ldots, N)$ に対する予測値を $\hat{y}_n$, 実測値を y_n とする

平均二乗誤差の平方根 (Root Mean Squared Error: RMSE)

$$\mathrm{RMSE} = \left(\frac{1}{N} \sum_n (y_n - \hat{y}_n)^2 \right)^{1/2}$$

平均絶対値誤差 (Mean Absolute Error: MAE)

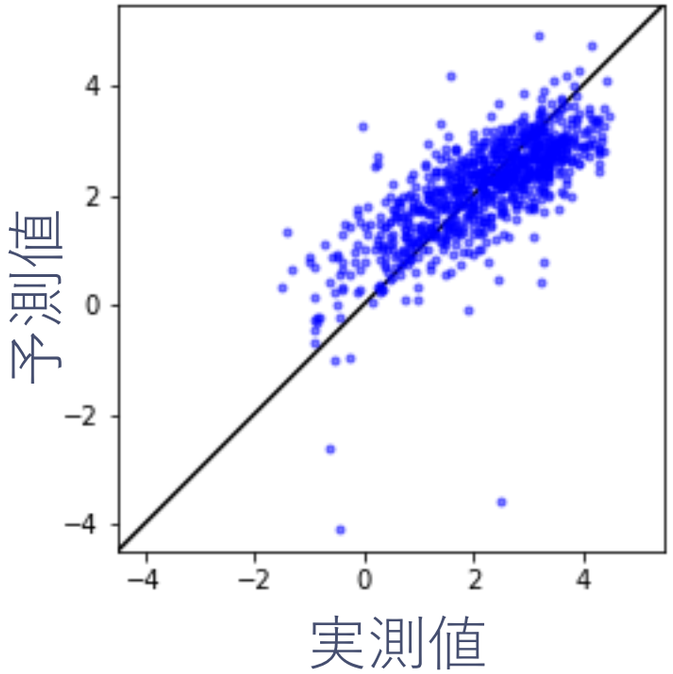
$$\mathrm{MAE} = \frac{1}{N} \sum_n |y_n - \hat{y}_n|$$

- モデルの精度を表す指標
- 0以上の値をとり, 0に近いほどモデルの精度がよい

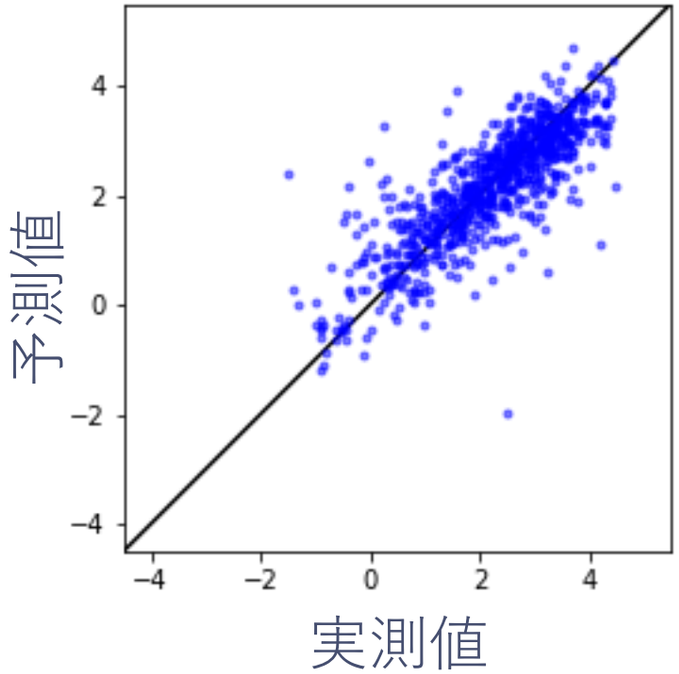
結果とモデルの評価

	R^2	RMSE	MAE
LASSO	0.5389	0.8145	0.6234
SVR	**0.6795**	**0.6790**	**0.4850**
RF	0.6546	0.7049	0.5193

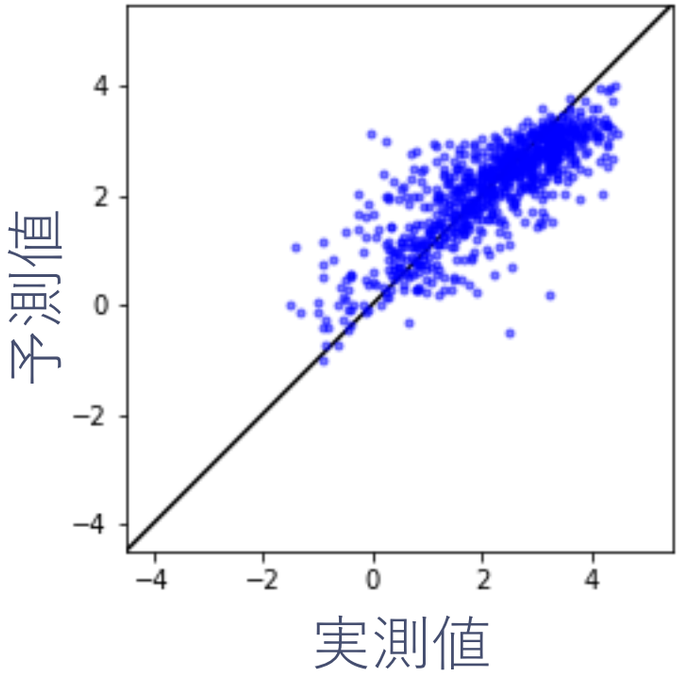
LASSO



SVR

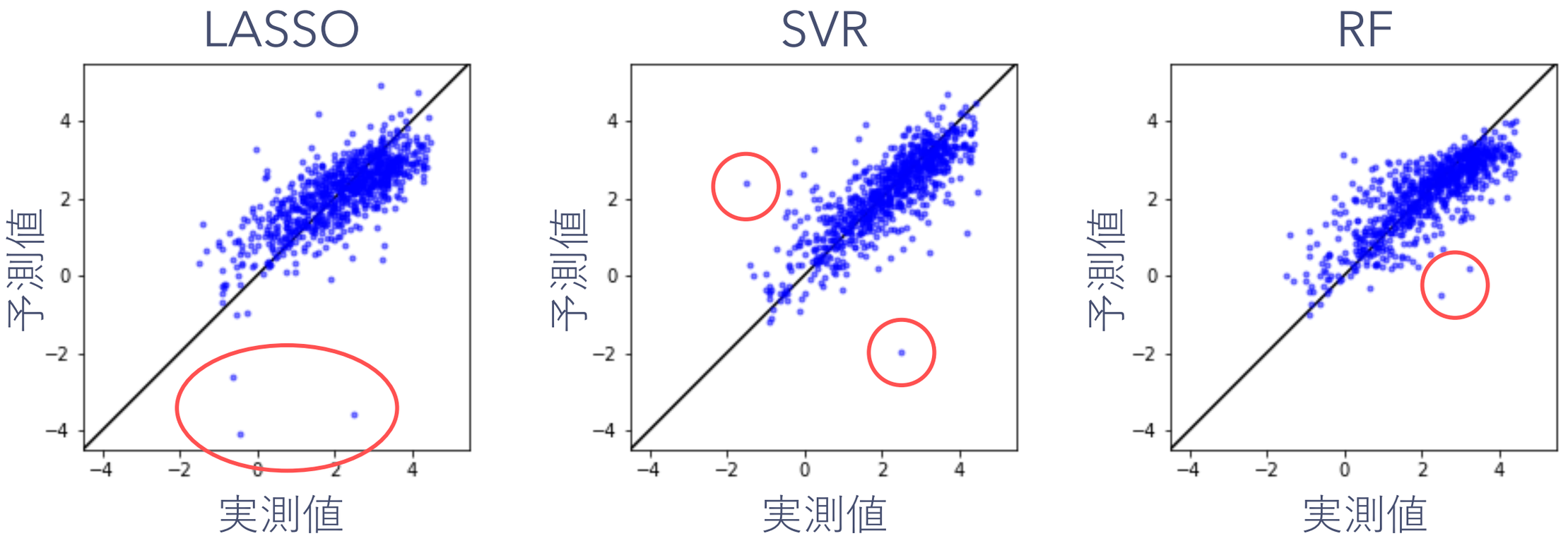


RF



➔ SVRの精度が最も良かった

結果を解釈する

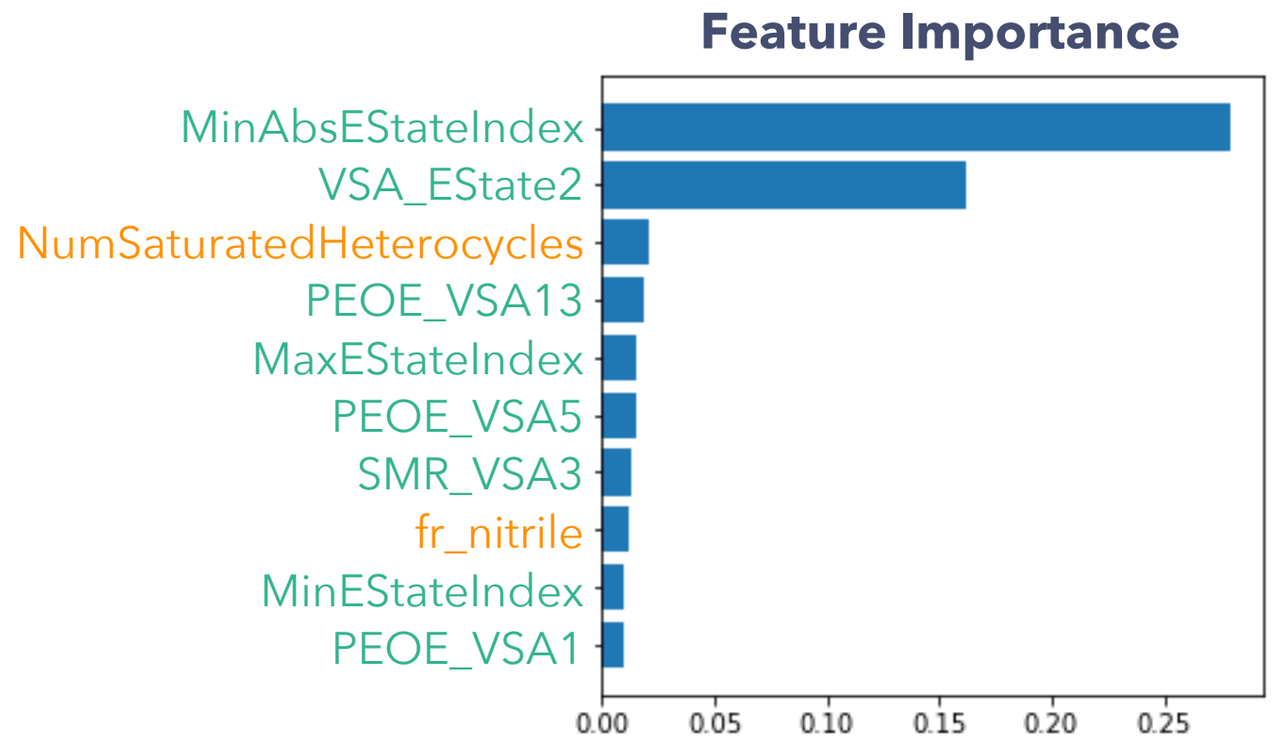


たとえば

- 誤差が大きいものを確認
- 記述子の重要度を確認

など

➔ 化学的知見, モデルの改善, ...



(電子状態　フラグメント)

まとめ

データ解析の流れ

1. データの前処理
2. モデルとその学習
3. モデルの評価
4. 次の解析など