イメージングXAFSデータ解析への機械学習的アプローチ

JASRI 産業利用推進室　高垣昌史

2019年9月4日(水)

SPring-8データ科学研究会 (第7回)
/ 第43回SPring-8先端利用技術ワークショップ
兵庫県マテリアルズ・インフォマティクス講演会（第3回）
「放射光計測インフォマティクス」

目次

イメージングXAFSとデータ解析のコンセプト

研究の背景

不均一試料における

化学状態の空間分布を可視化する

化学状態の推定 ← XAFSスペクトル

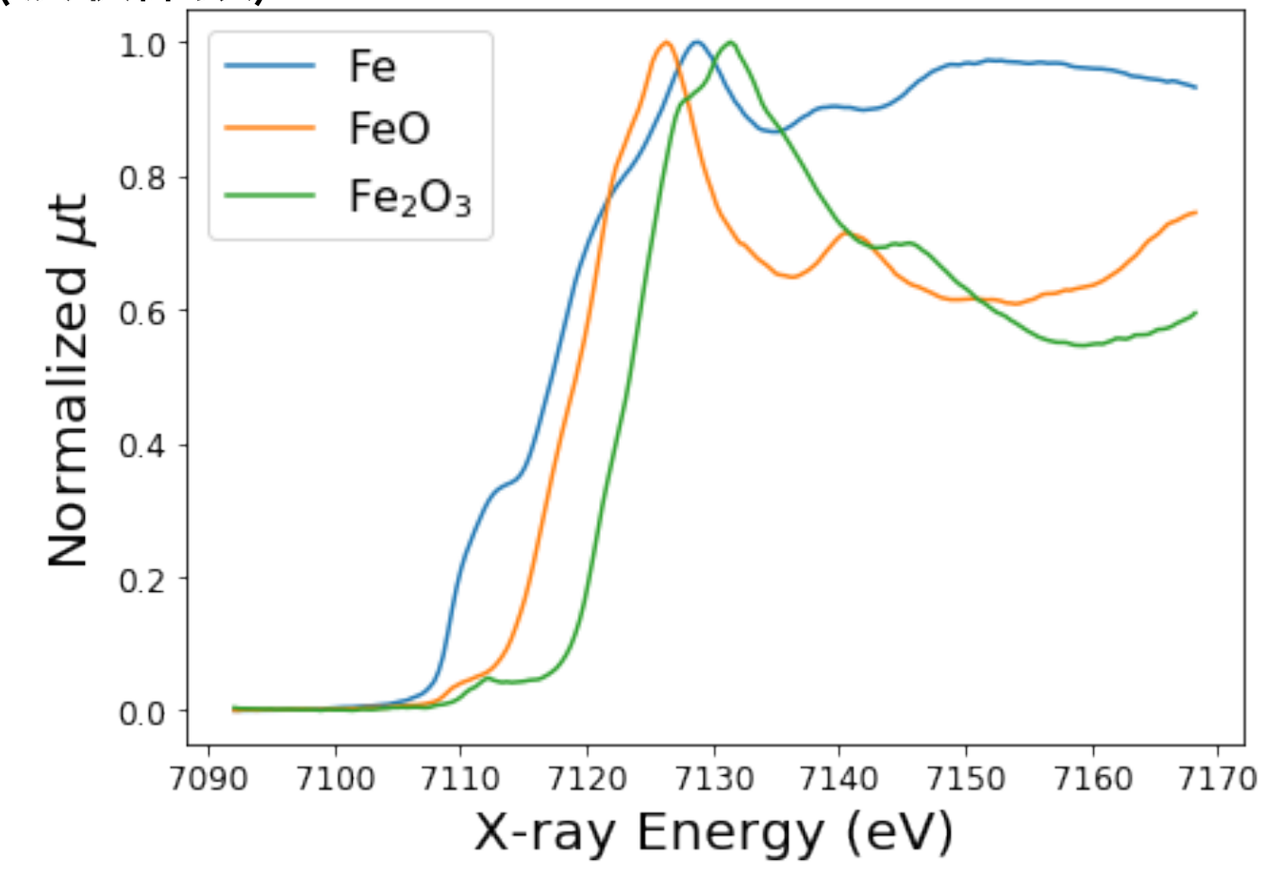
空間分布の同定 ← イメージング

イメージングXAFS

XAFSスペクトル

X-ray Absorption Fine Structure (X線吸収微細構造)

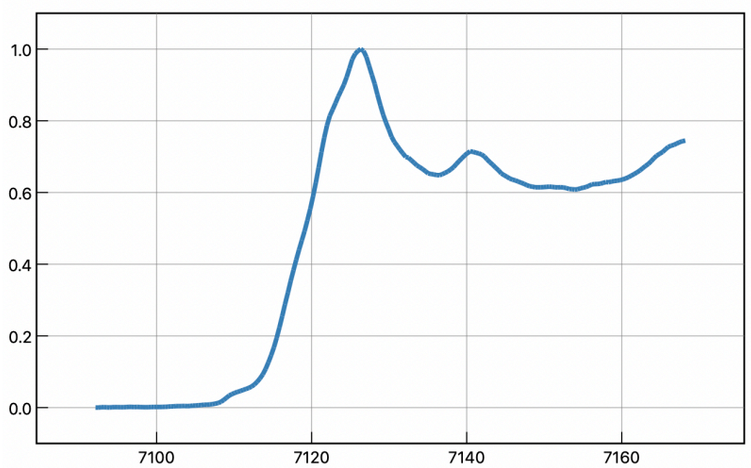
Fe-K 吸収端でのスペクトルの例



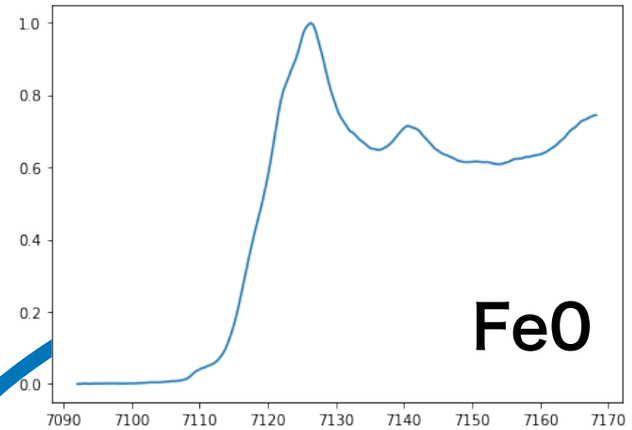
- 吸収端：元素固有の吸収エネルギー
- 注目元素の情報を選択的に取得できる
- 化学状態の情報が含まれる

XAFS指紋法

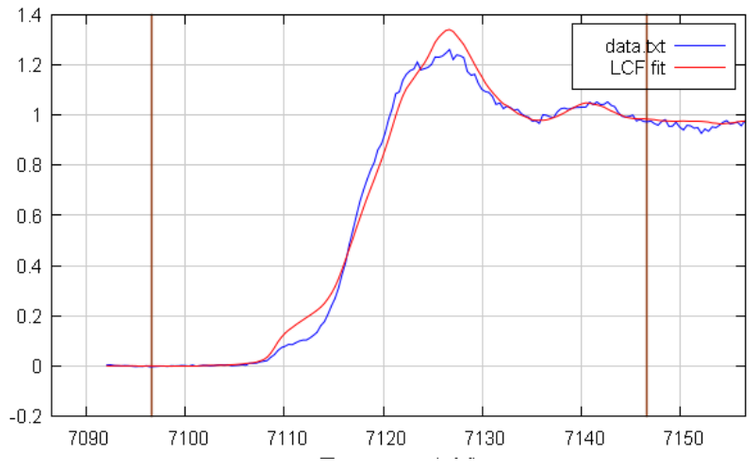
解析対象のスペクトル



標準試料のスペクトル

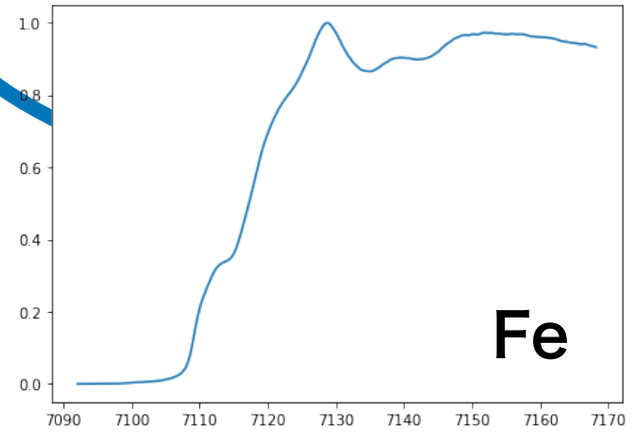


FeOと同定



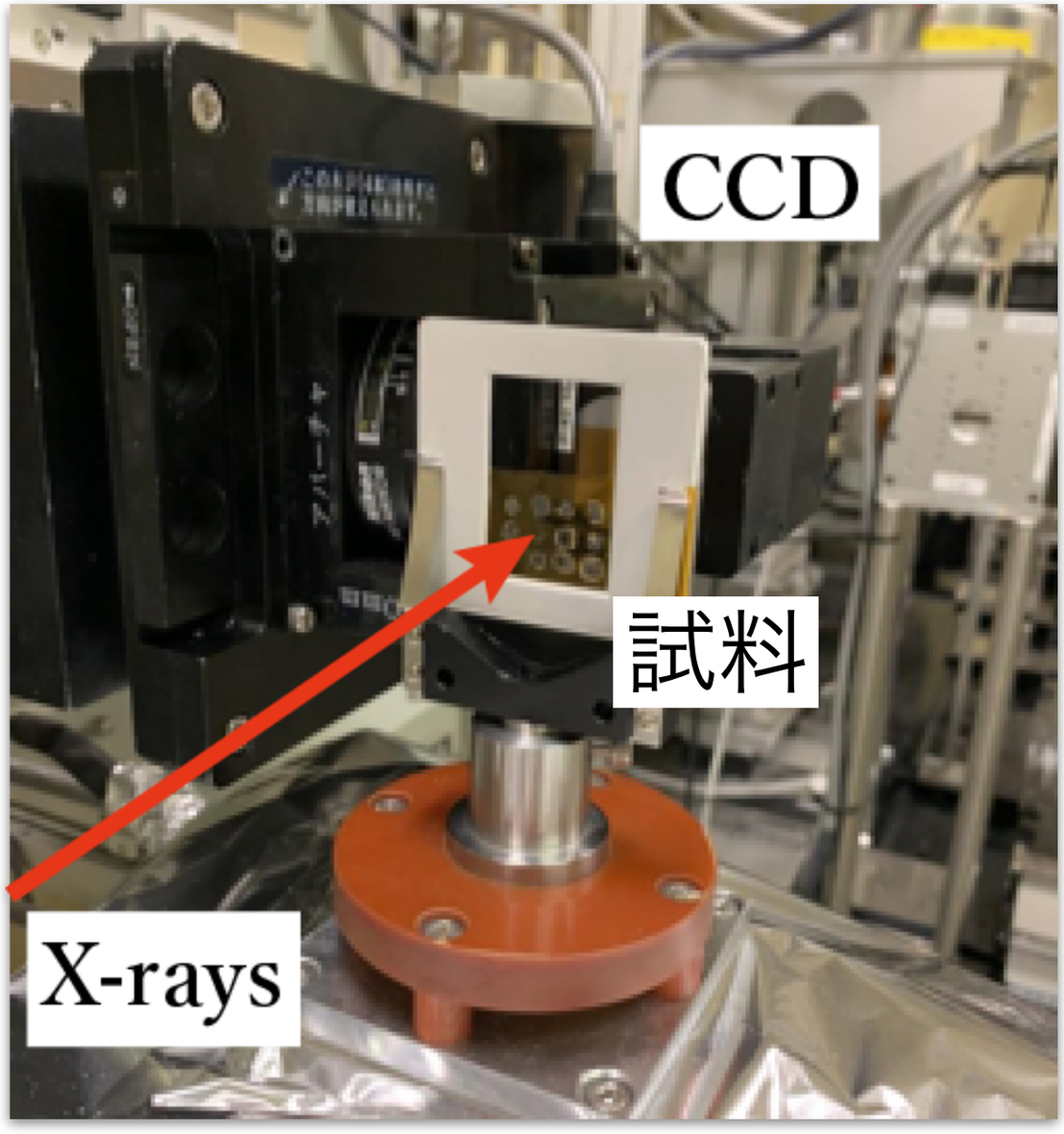
線形結合フィット

= 0.586 FeO
+ 0.414 Fe



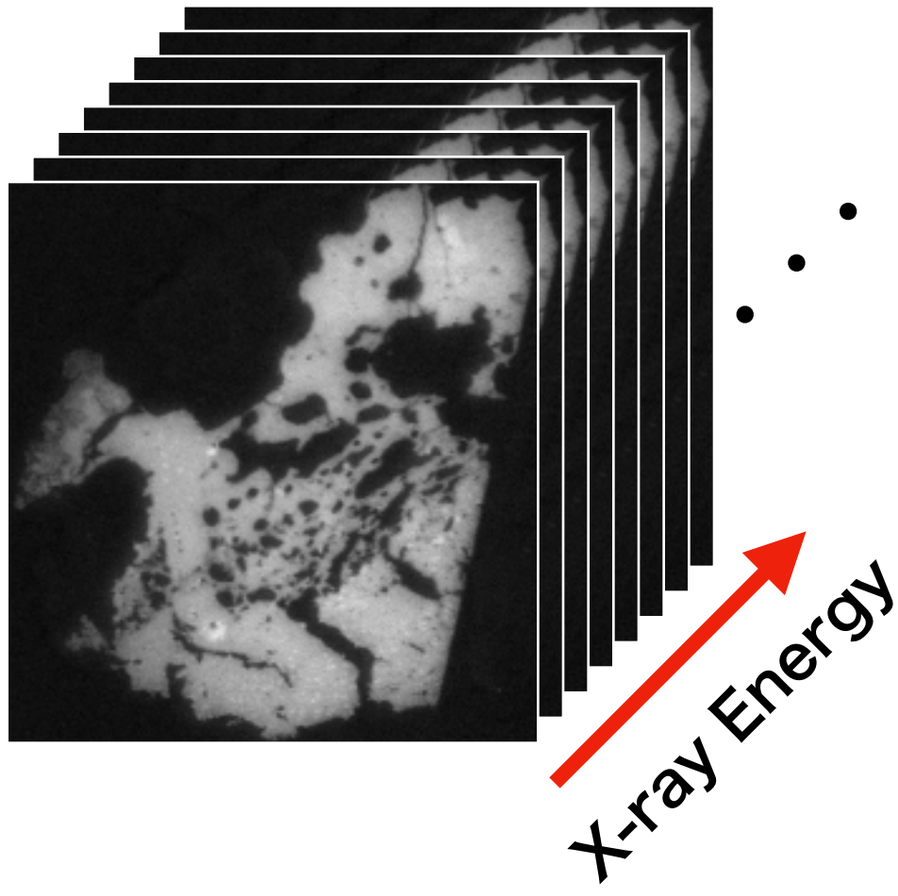
- 標準試料のスペクトルと**形状を比較**し、化学状態を推定する
- 解析者の**主観的判断**が色濃い解析手法

イメージングXAFS



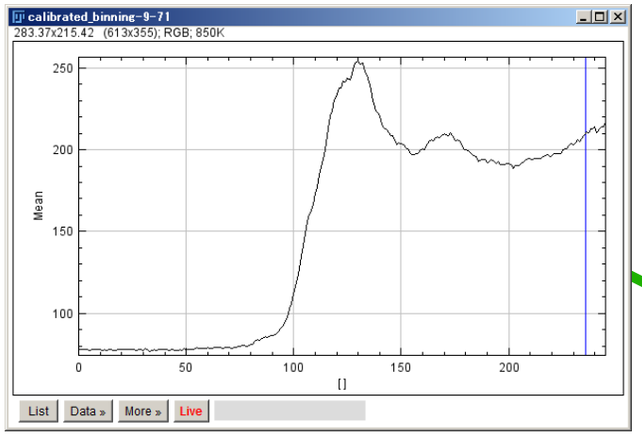
X線エネルギーを変化させつつ
透過像を複数取得

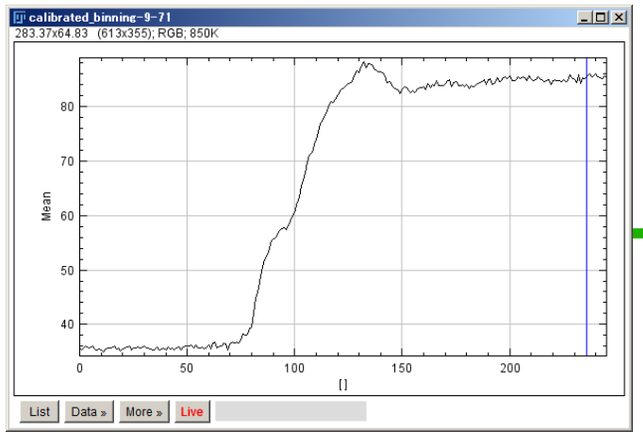
吸収係数像

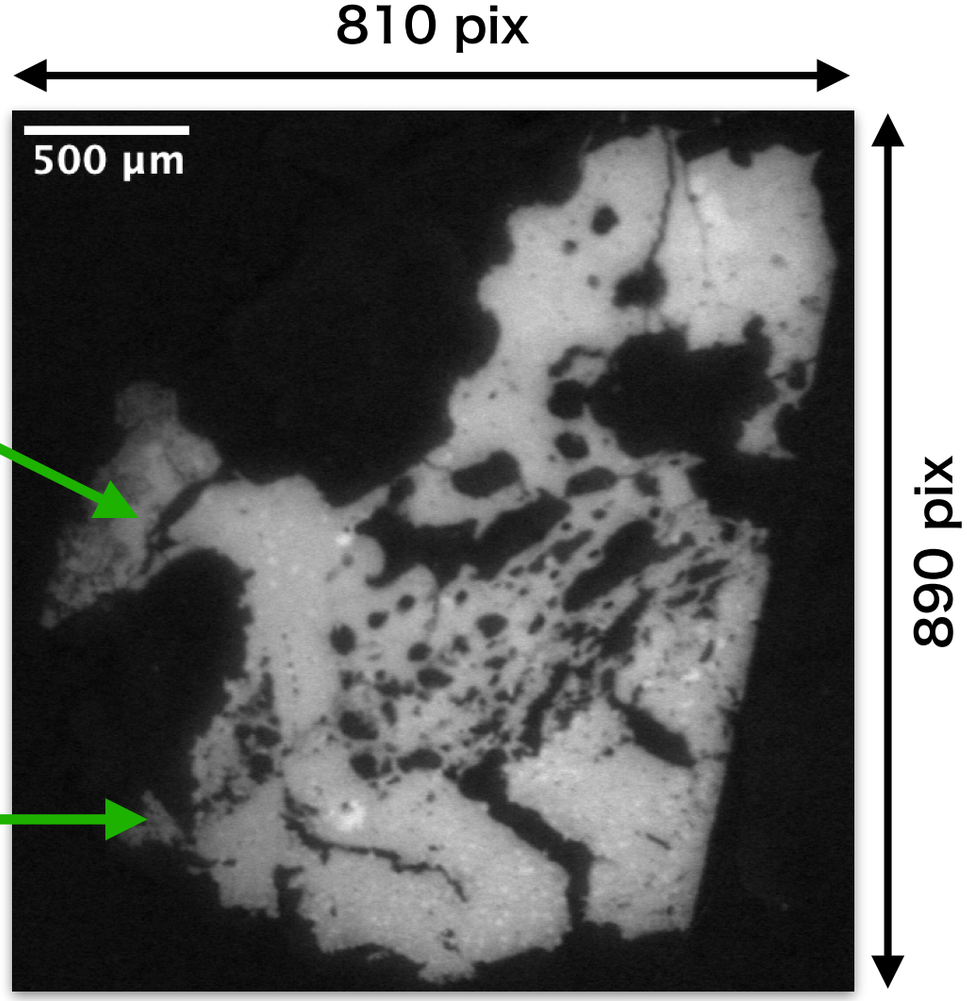


XAFSスペクトルの
空間分布が得られる
3次元データ

イメージングXAFSデータ







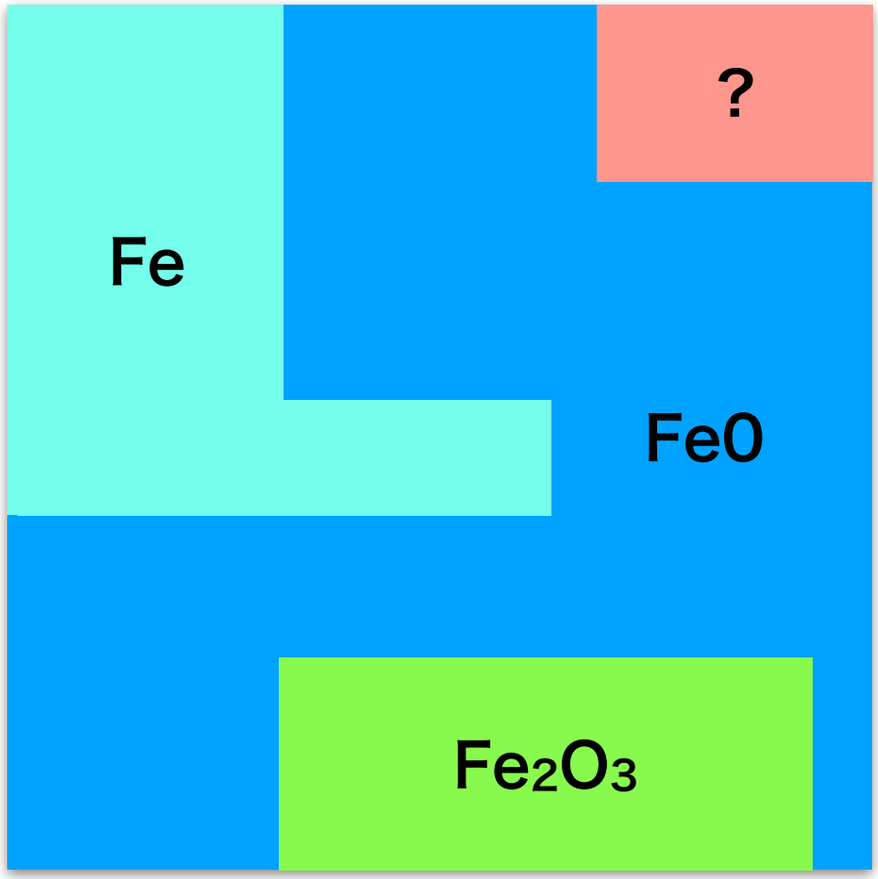
総ピクセル数: 720,900 pix

解析を困難にしている点

膨大な数のスペクトル ➡ **手作業での解析に限界**

多次元データ ➡ **全体像の把握が困難**

研究の目的



解析結果の概念図
化学状態で分類, 領域の塗り分け

- 「人間的で曖昧な判断」の指紋法を
  機械学習で実現, 化学状態で分類

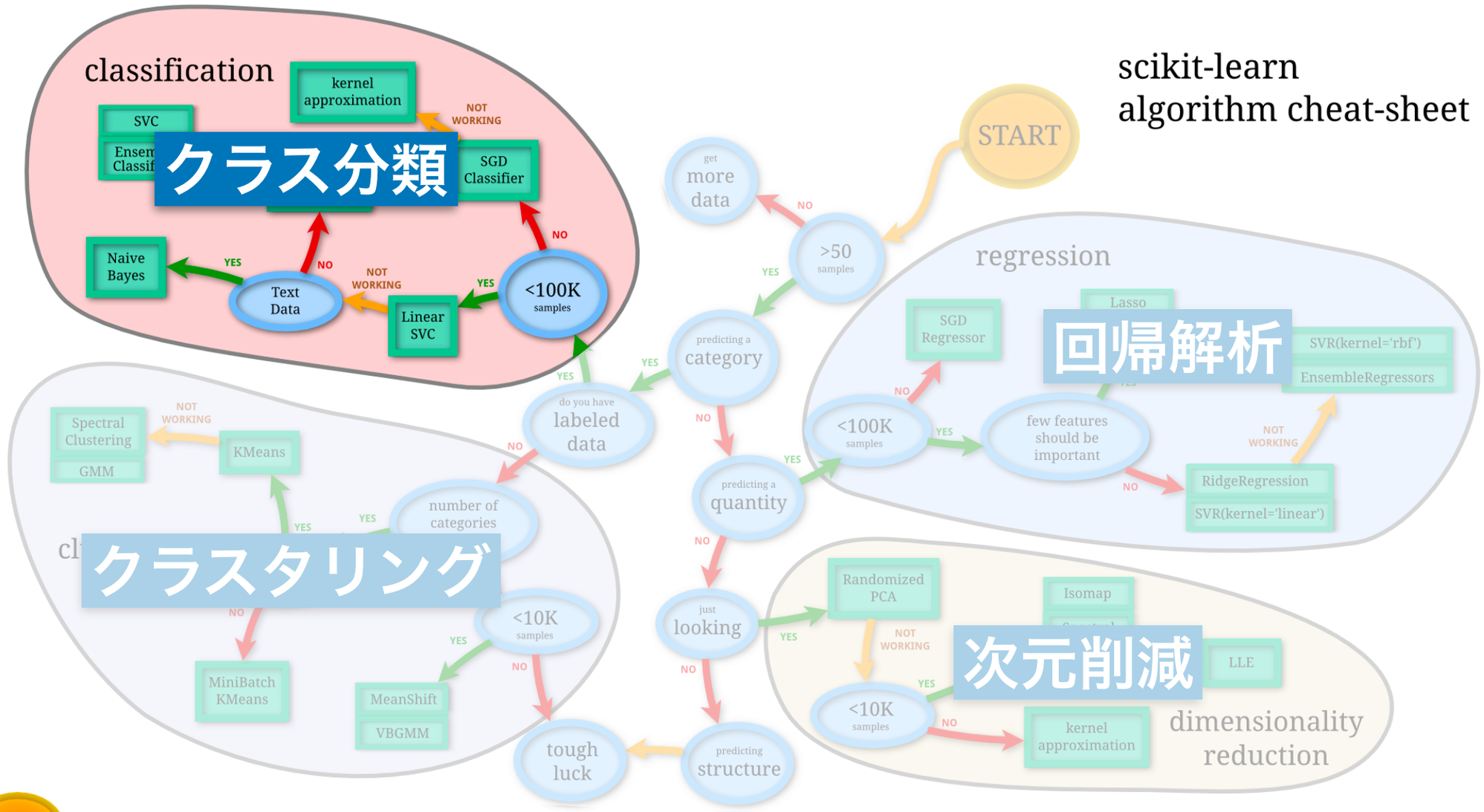
  スペクトル(1D配列) → 化学状態のラベル
- **次元を削減**し化学状態を可視化

- 全体像の把握 > 厳密性

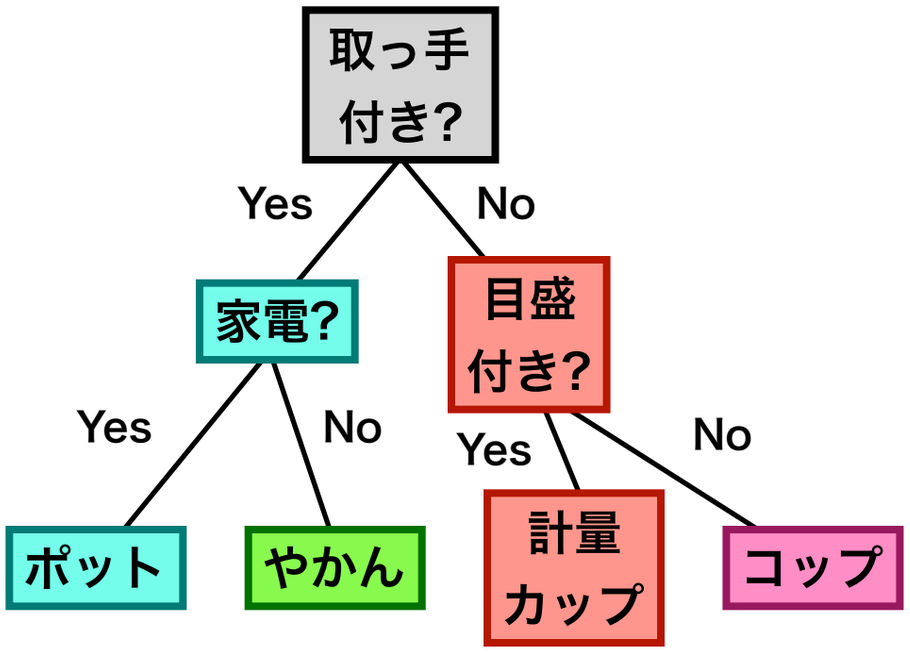
  (一目見て理解できる分類数に抑える)
  ～15

- 未知の要素の存在を示唆

機械学習アルゴリズム


scikit-learn
algorithm cheat-sheet
START
classification
クラス分類
kernel approximation
NOT WORKING
SVC
Ensem
Classif
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
YES
<100K samples
Linear SVC
get more data
NO
>50 samples
YES
regression
回帰解析
Lasso
SGD Regressor
SVR(kernel='rbf')
EnsembleRegressors
predicting a category
YES
do you have labeled data
NO
<100K samples
YES
few features should be important
NOT WORKING
RidgeRegression
SVR(kernel='linear')
NO
predicting a quantity
YES
NO
Spectral Clustering
NOT WORKING
GMM
KMeans
YES
number of categories
YES
クラスタリング
NO
<10K samples
YES
NO
MiniBatch KMeans
MeanShift
VBGMM
tough luck
just looking
NO
predicting structure
Randomized PCA
YES
NOT WORKING
Isomap
次元削減
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
Back
scikit learn

RandomForest



キッチン用品を区別する決定木

- 教師あり学習
- 複数の**決定木**によるクラス分類
- 教師データとして**クラス定義**を与える
- 「**どのクラスに似ているか**」を判断する
  (≠フィッティング)
- scikit-learn : **RandomForestClassifier**

与えるべきデータ構造は?

RandomForestClassifier の受け付けるデータフォーマット

特徴量

特徴量行列

(2D配列)

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{pmatrix}$$

サンプル

データ数 = M

特徴量数 = N

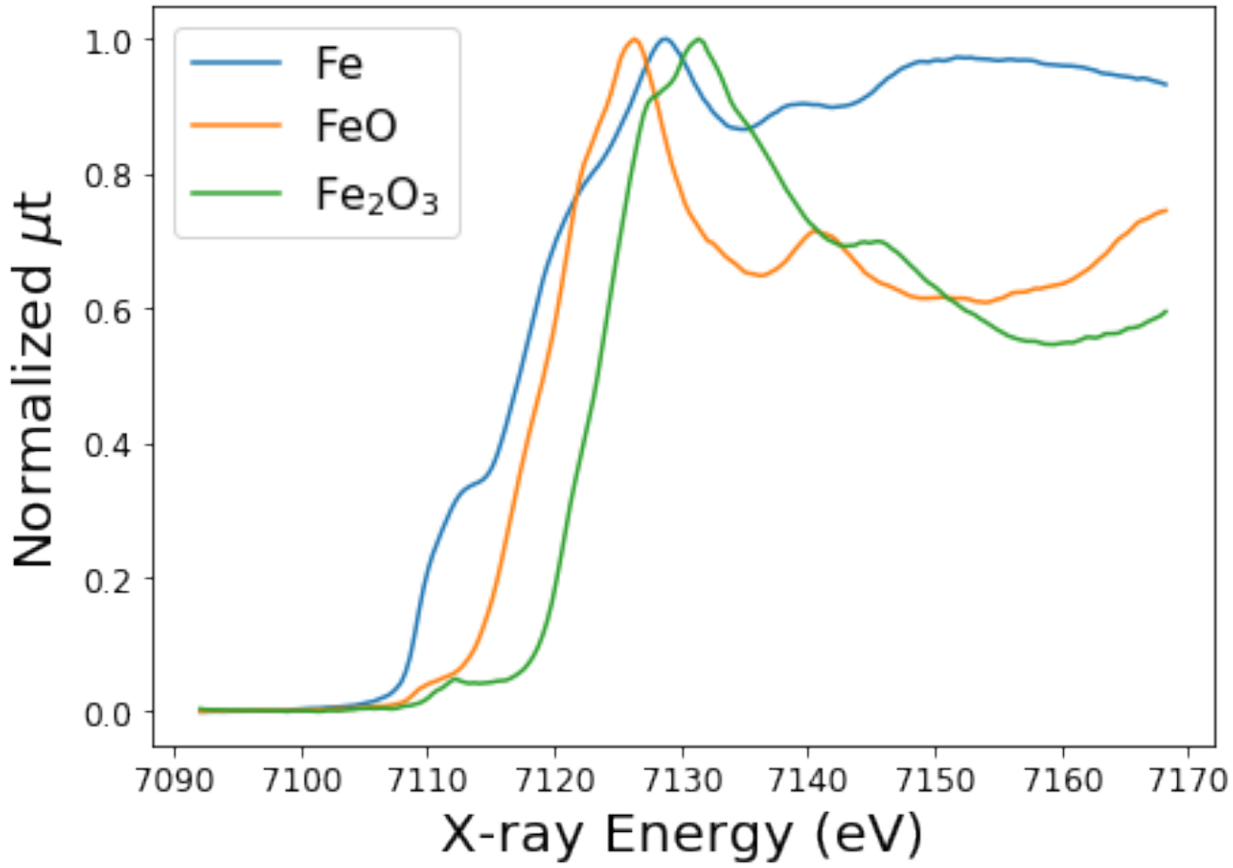
目的変数ベクトル

(1D配列)

$$\boldsymbol{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_M \end{pmatrix}$$

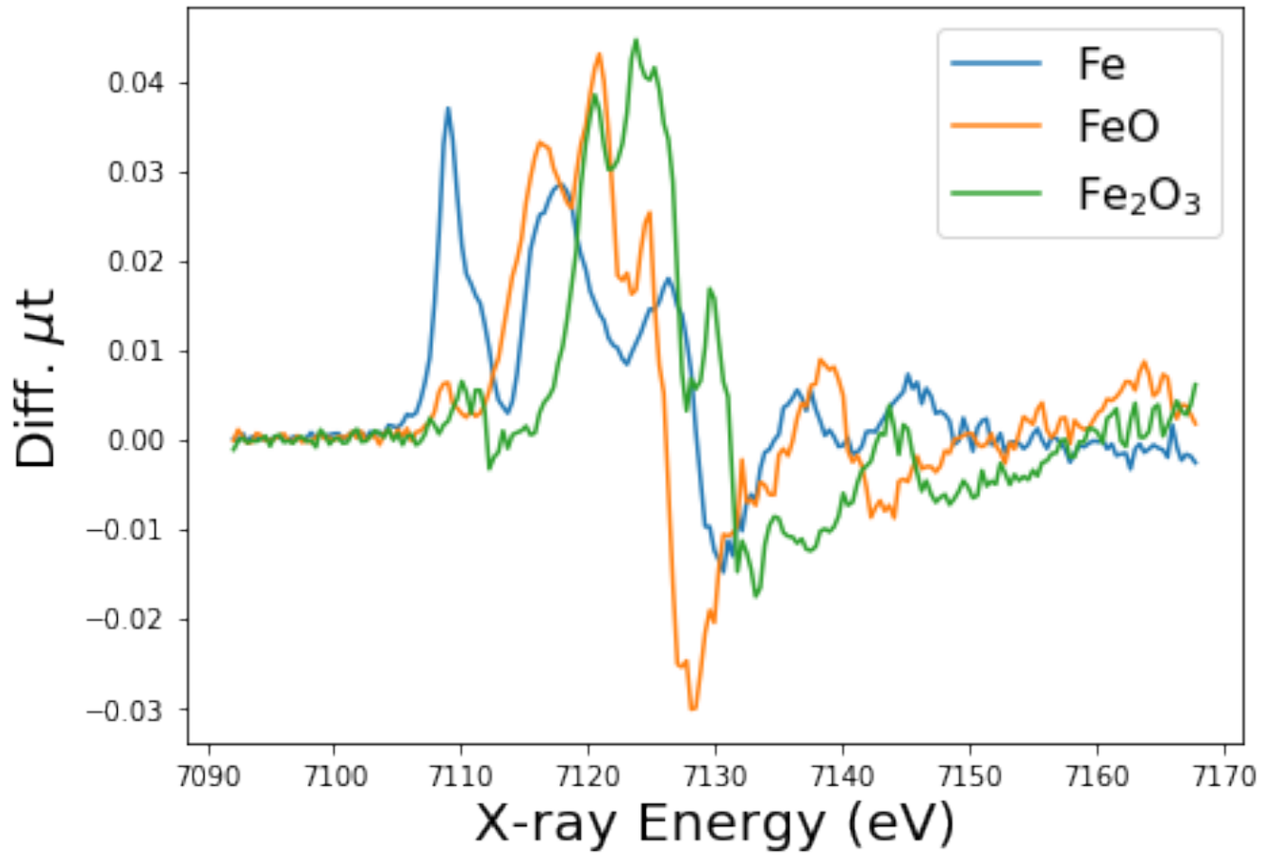
何を「特徴量」「目的変数」とするか?

特徴量をどう定義する?



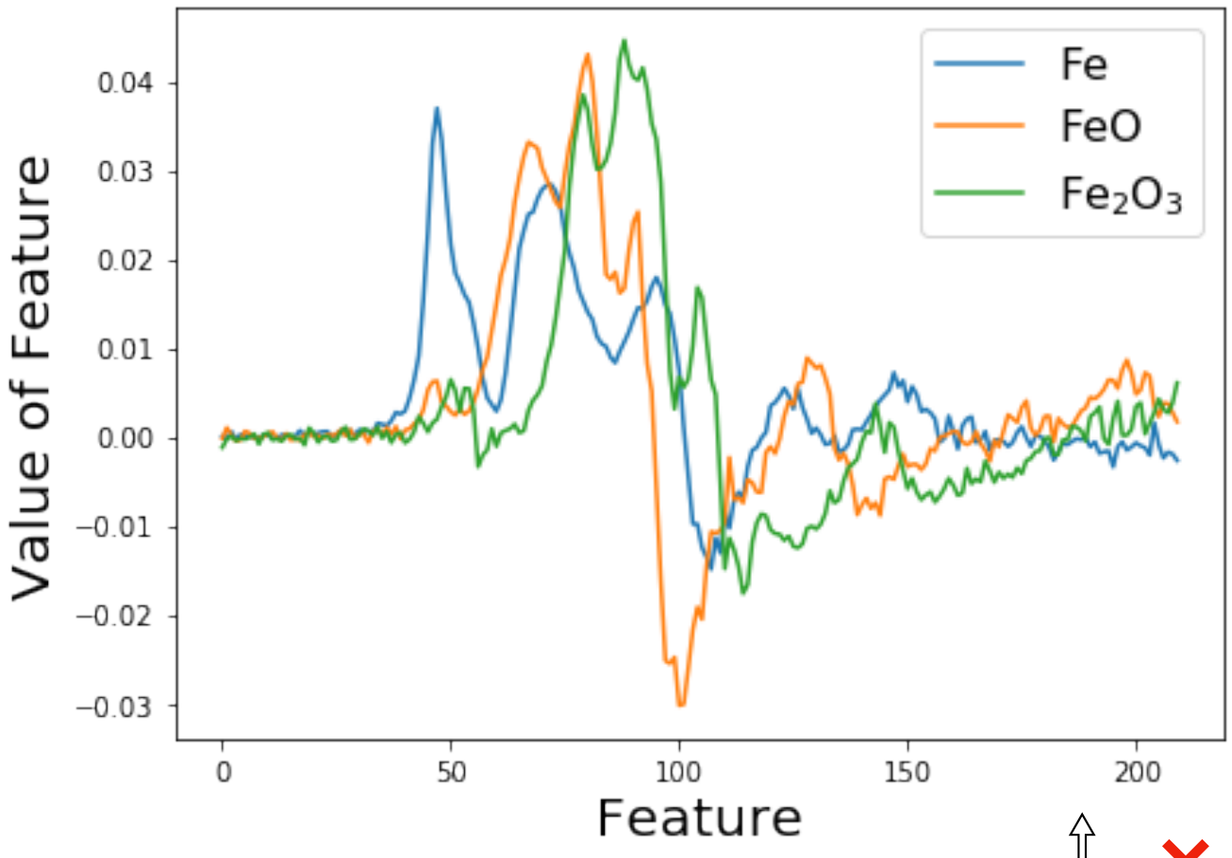
全てのエネルギー点 (210点) における信号強度

形状変化を評価⇨微分強度



全てのエネルギー点 (210点) における信号強度

特徴量の定義

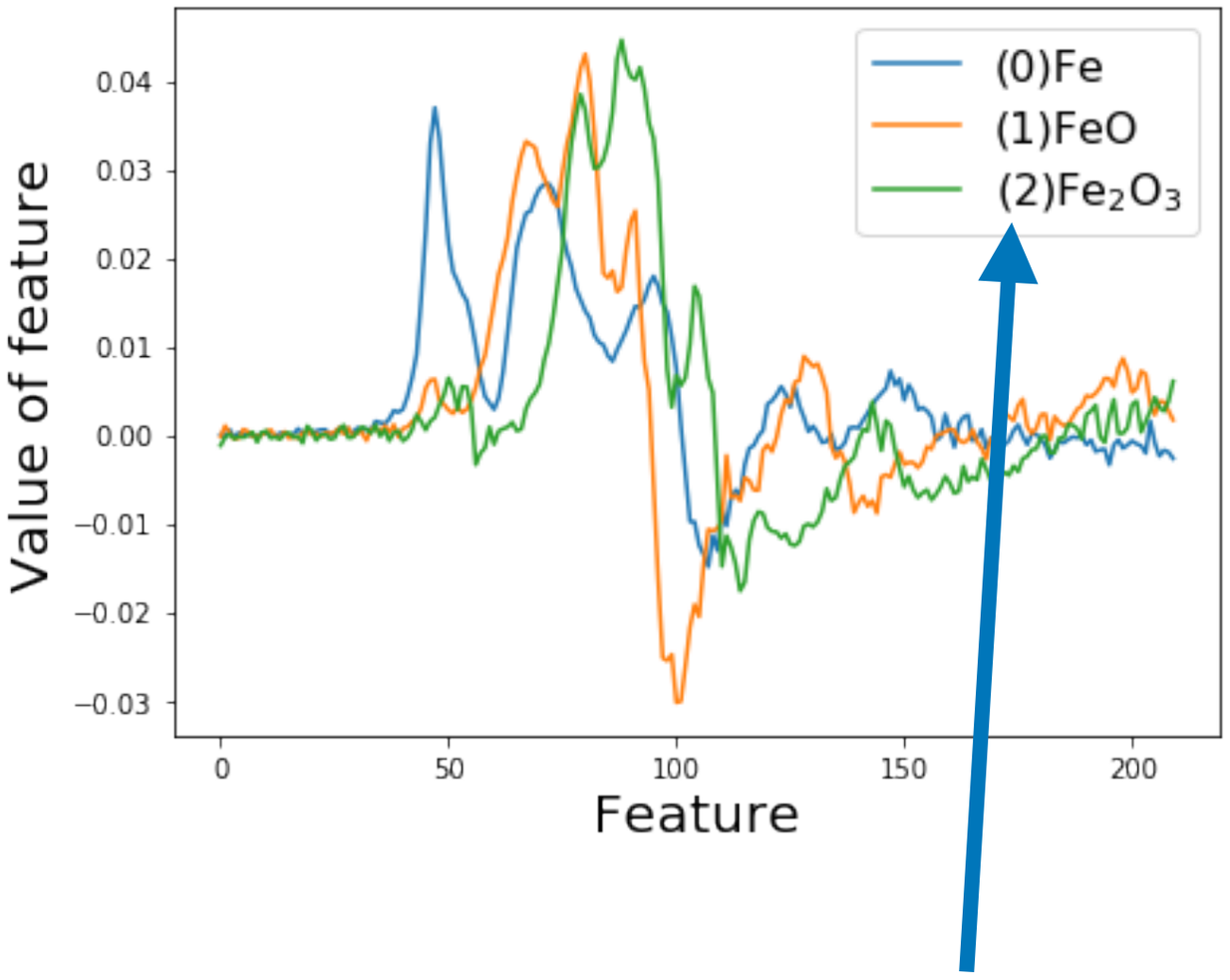


⇧ × エネルギー

◎ 通し番号

計算機から見れば「スペクトルはただの1次元配列」

目的変数をどう定義する?



各スペクトルの通し番号

＝ クラス名

RandomForestに与えるデータ構造

微分スペクトルデータそのもの

特徴量行列

(2D配列)

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & & x_{2N} \\ \vdots & & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{pmatrix}$$

$(N = 210)$

サンプル

目的変数ベクトル

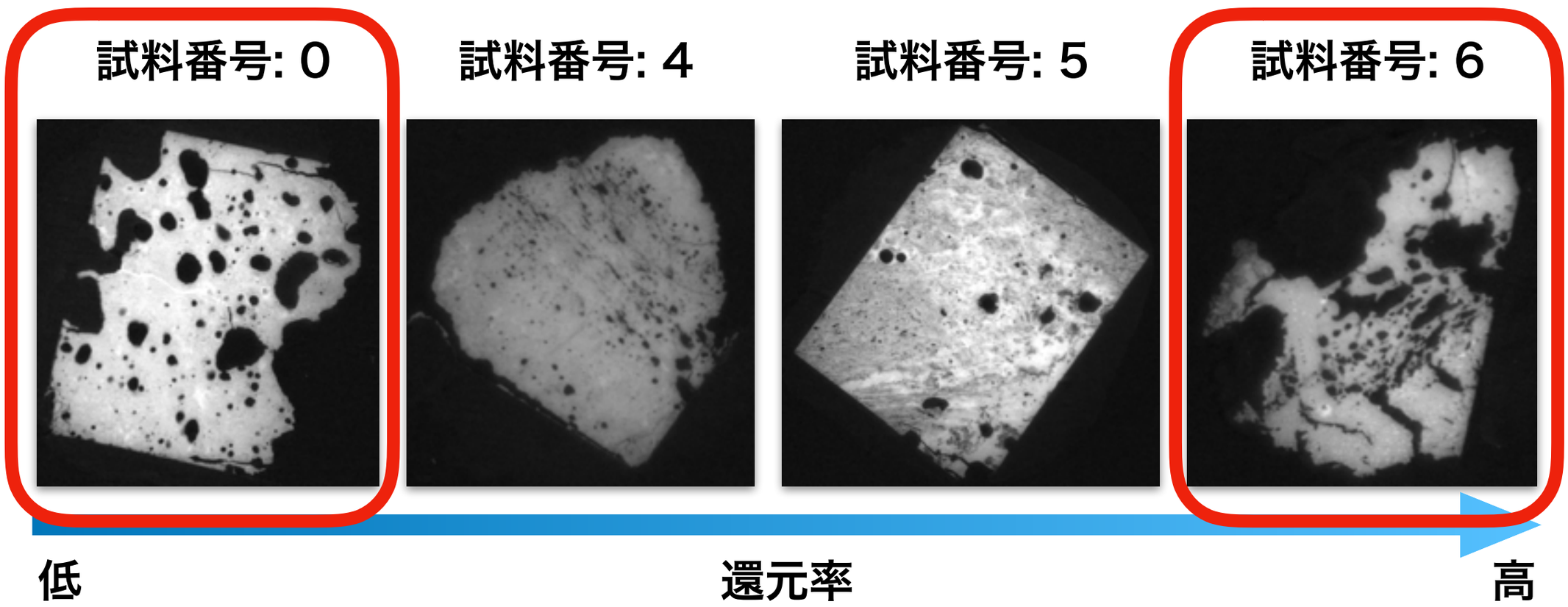
(1D配列)

$$\boldsymbol{y} = \begin{pmatrix} 0 & 1 & \cdots & M \end{pmatrix}$$

クラス名 (通し番号)

測定試料と測定条件

焼結鉱

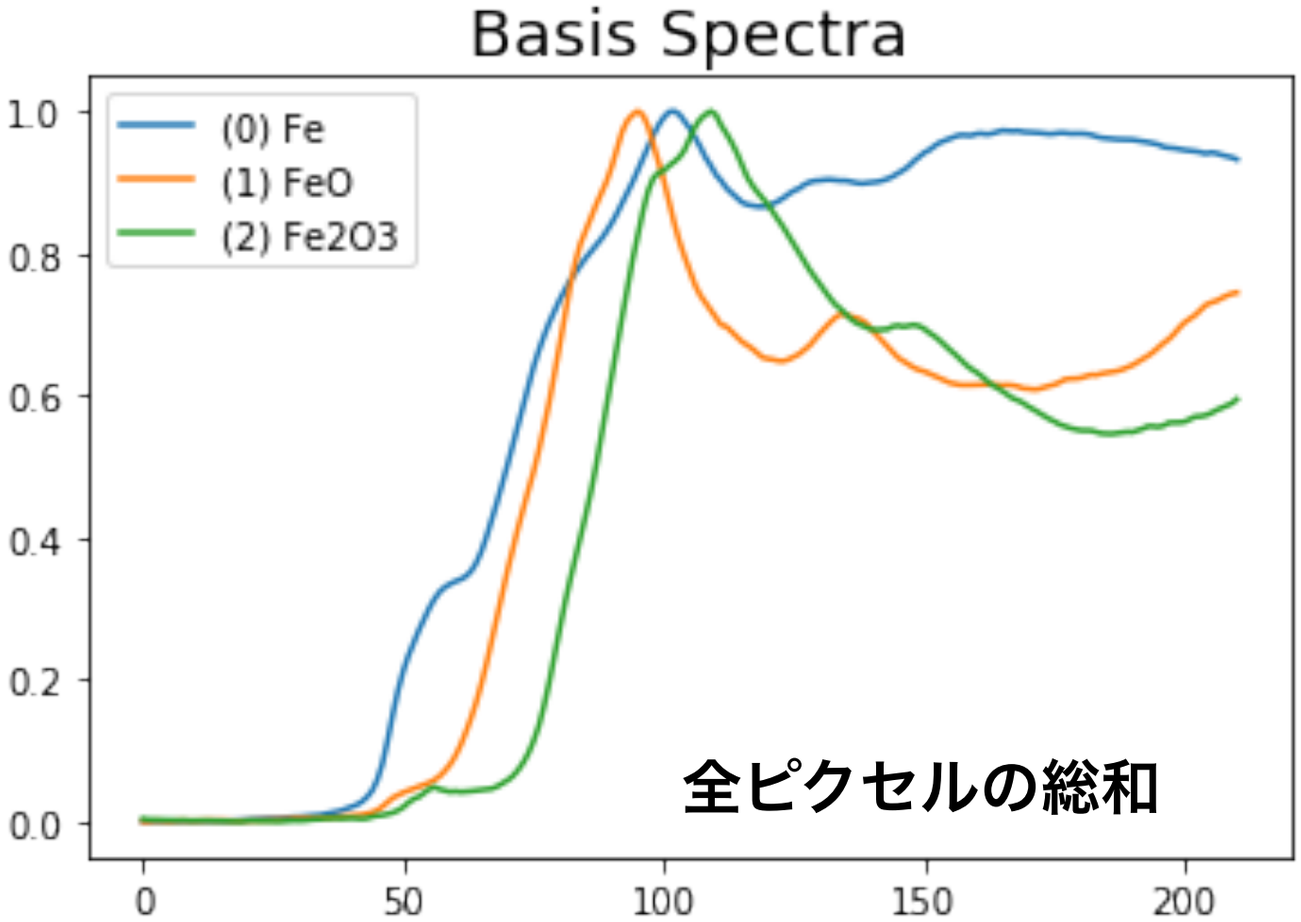


- 鉄の精錬工程における中間材料
- **Fe, Fe酸化物**等が主体
- Ca, Si, Al 等 鉱滓(こうさい)(スラグ)
- 透過測定用に薄片化 (～10 μm)

測定条件

- SPring-8 産業利用ビームライン BL14B2
- 透過配置: Fe-K 吸収端, Si(111)
- 測定時間: 100 分/測定 (露光時間: 2 秒/エネルギー)
- CCDカメラ
  - 撮像視野: (w)11mm × (h)5 mm
  - ピクセルサイズ: 2.92 μm/pix
- データサイズ: 7 GB/測定

標準試料スペクトル



0. Fe-foil

ウスタイト
1. FeO (BN希釈ペレット)

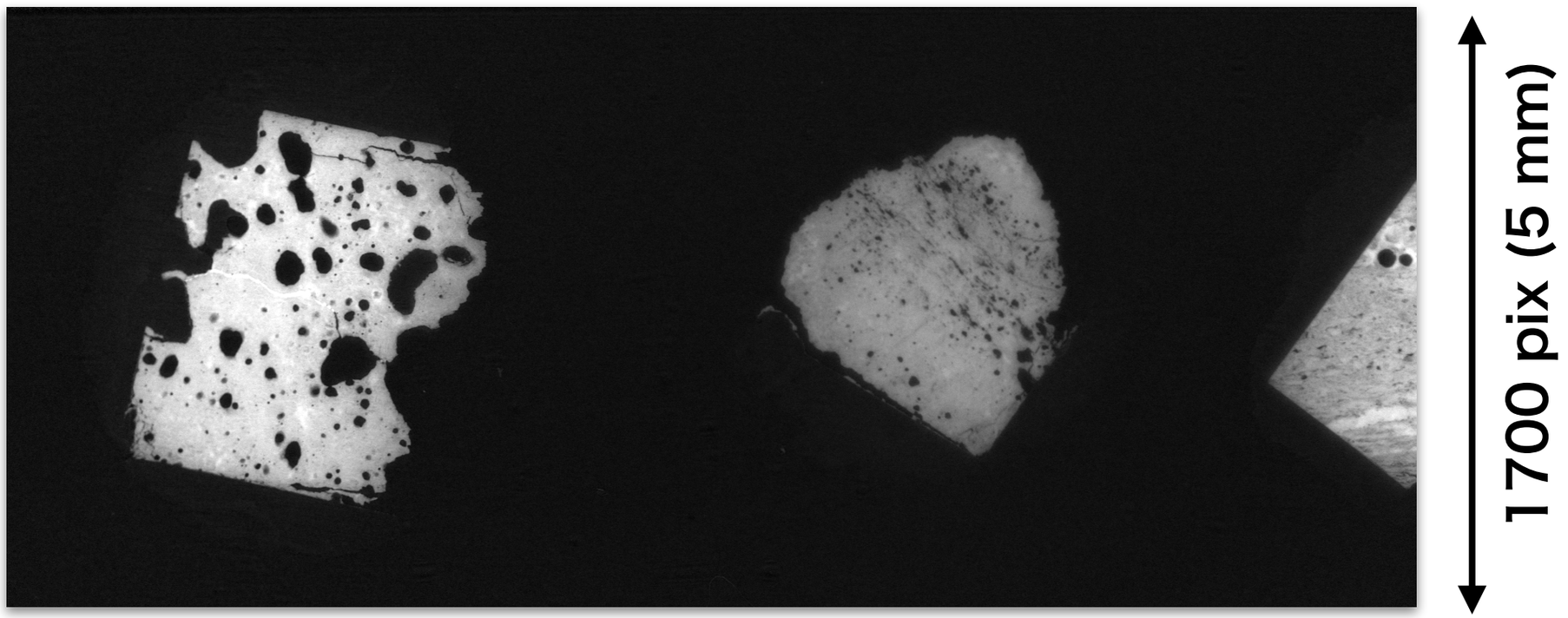
ヘマタイト
2. Fe_2O_3 (BN希釈ペレット)

微分スペクトルを学習データとして利用

測定データの事前調査と前処理

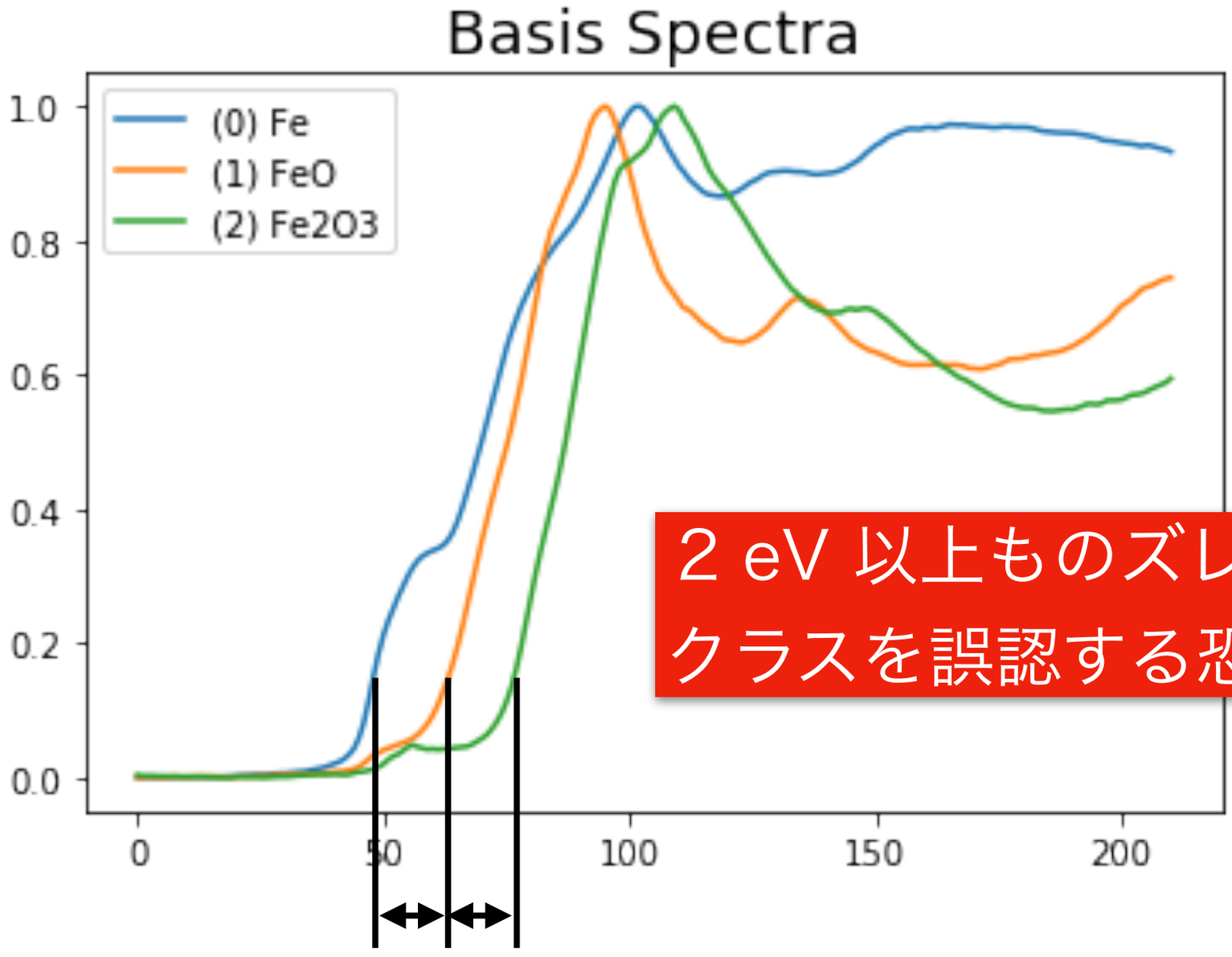
発見された問題点①

エネルギー校正の必要性



- **入射X線エネルギーに空間分布が確認された**
- **最大 2 eV 以上のズレ**

標準試料スペクトル



2 eV 以上ものズレがあると
クラスを誤認する恐れがある

～ 4eV

発見された問題点②

データサイズ

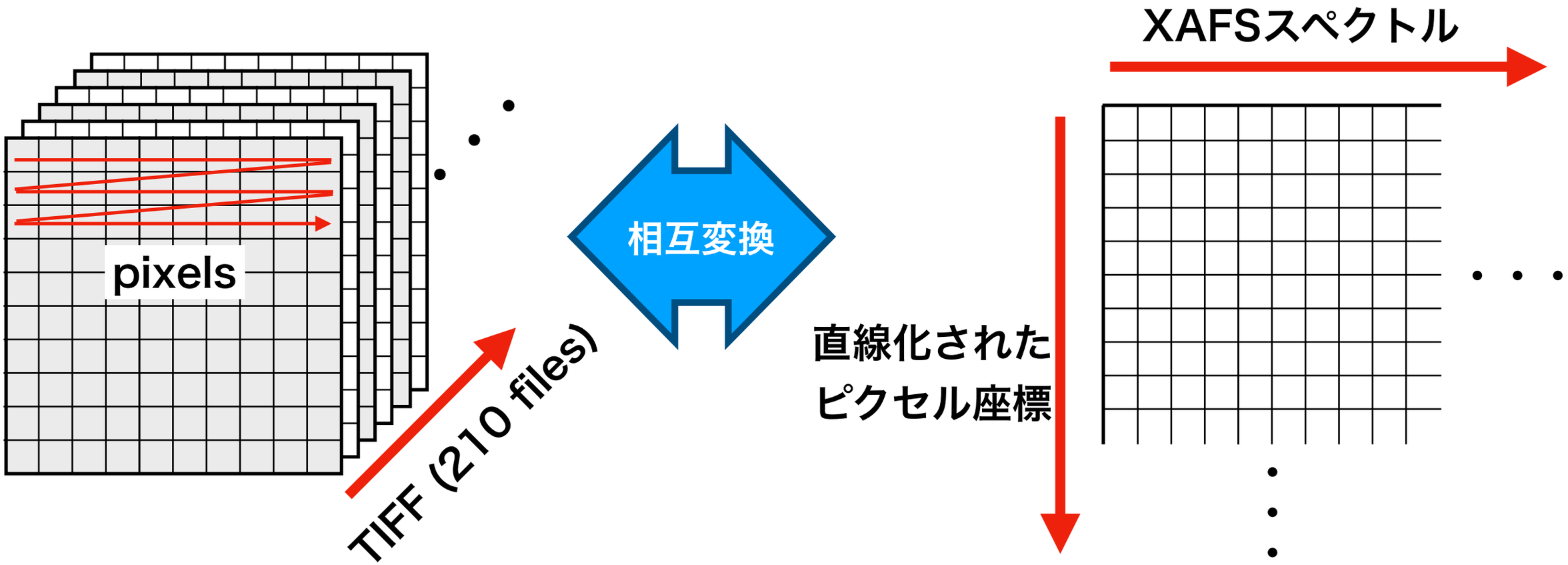
特徴量行列
(2D配列)

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{pmatrix}$$

- データ総量: **7 GB**
- 画像フォーマット: TIFF (小数を扱うため)
- libtiff (標準的TIFFライブラリ)の上限: **4 GB/ファイル**
- **データの一部領域を切り出す必要性**

その他の前処理

データ形式変換



測定データ形式
(3D 配列)

機械学習のデータ形式
(2D 配列)

作成した前処理プログラム

- エネルギー校正

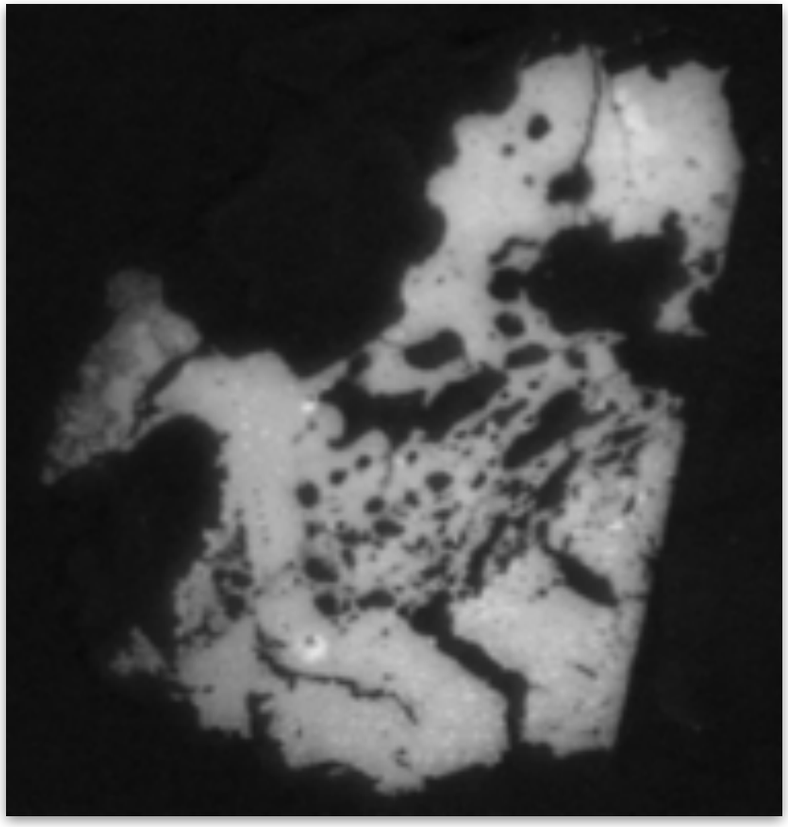
  ズレ量測定/マップ生成、マップに基づく校正

- 領域切り出し
- データ形式相互変換
- 規格化
- 隣接ピクセル平均化(S/N向上のため)
- バックグラウンド領域認識用マスク生成

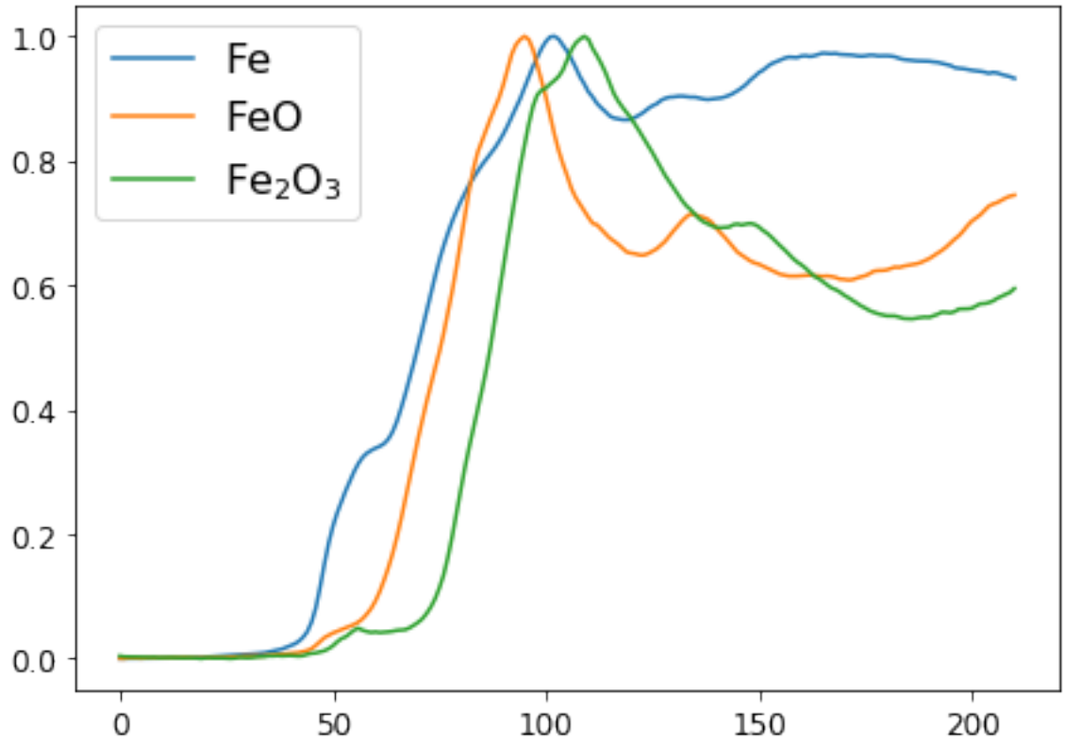
処理が重いので、いずれも C++ で作成

RandomForest による化学状態分類と可視化

試料番号:6



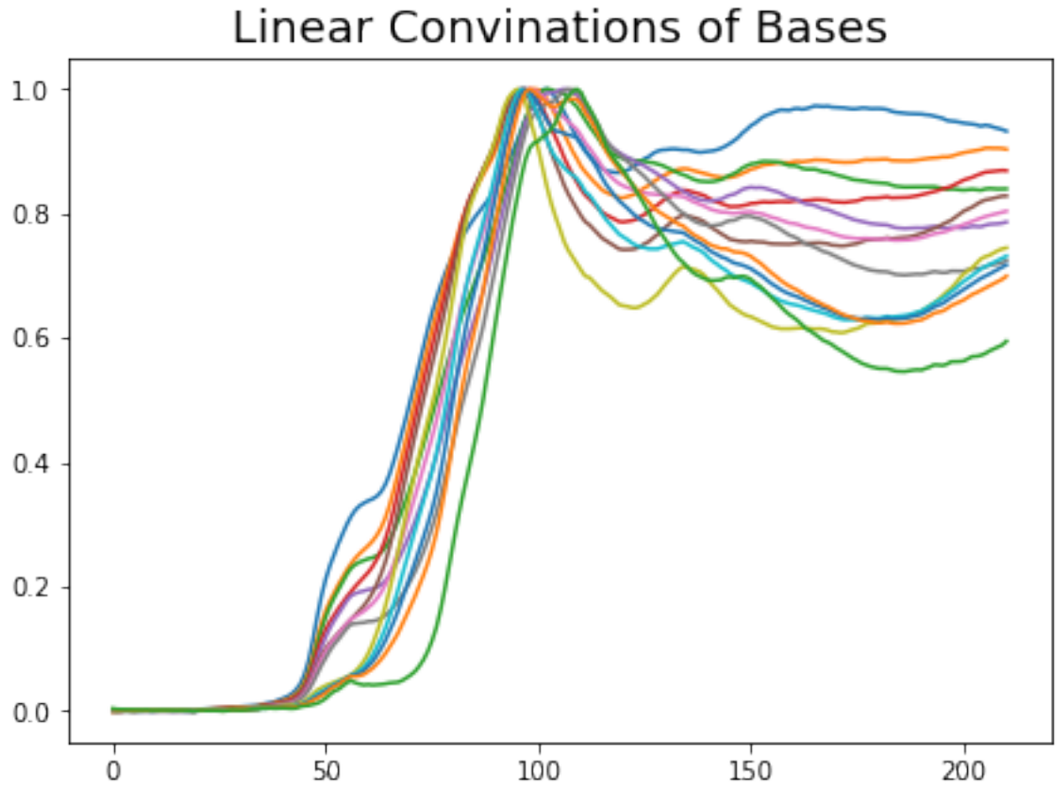
最も還元率が高い



基底スペクトルを
Fe, FeO, Fe_2O_3 と想定

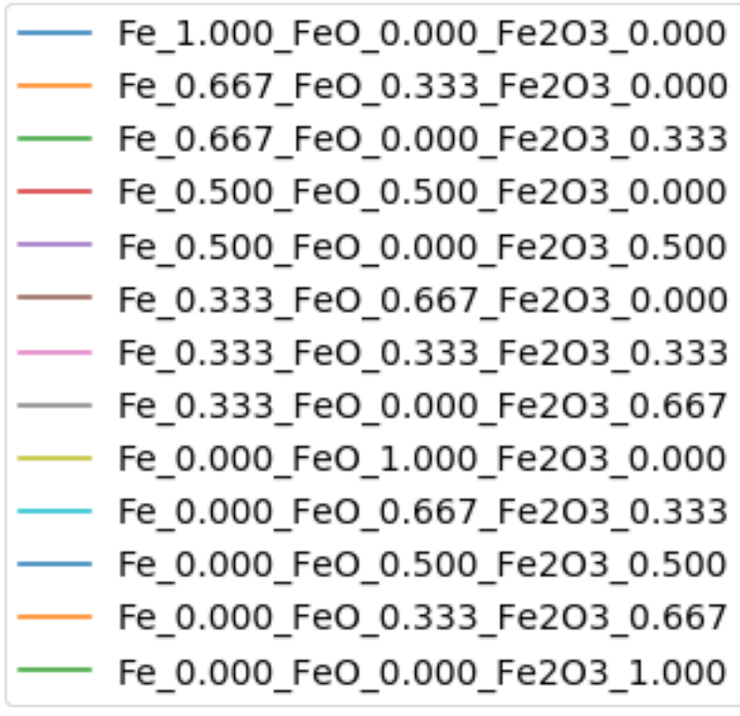
基底スペクトルの線型結合

バリエーション = 13



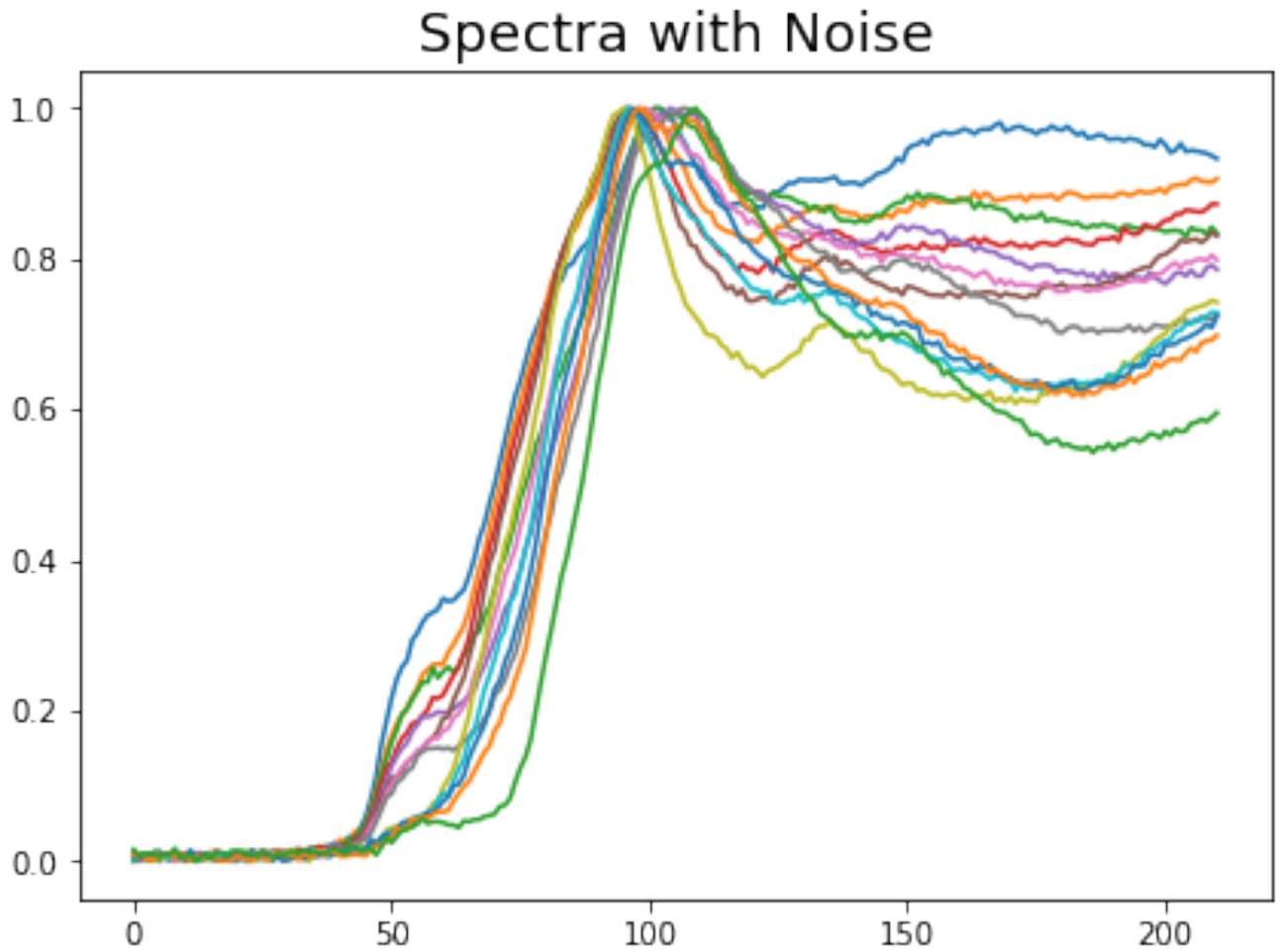
上から順に 0, 1, 2 … (クラス)

クラスのラベル(混合比)



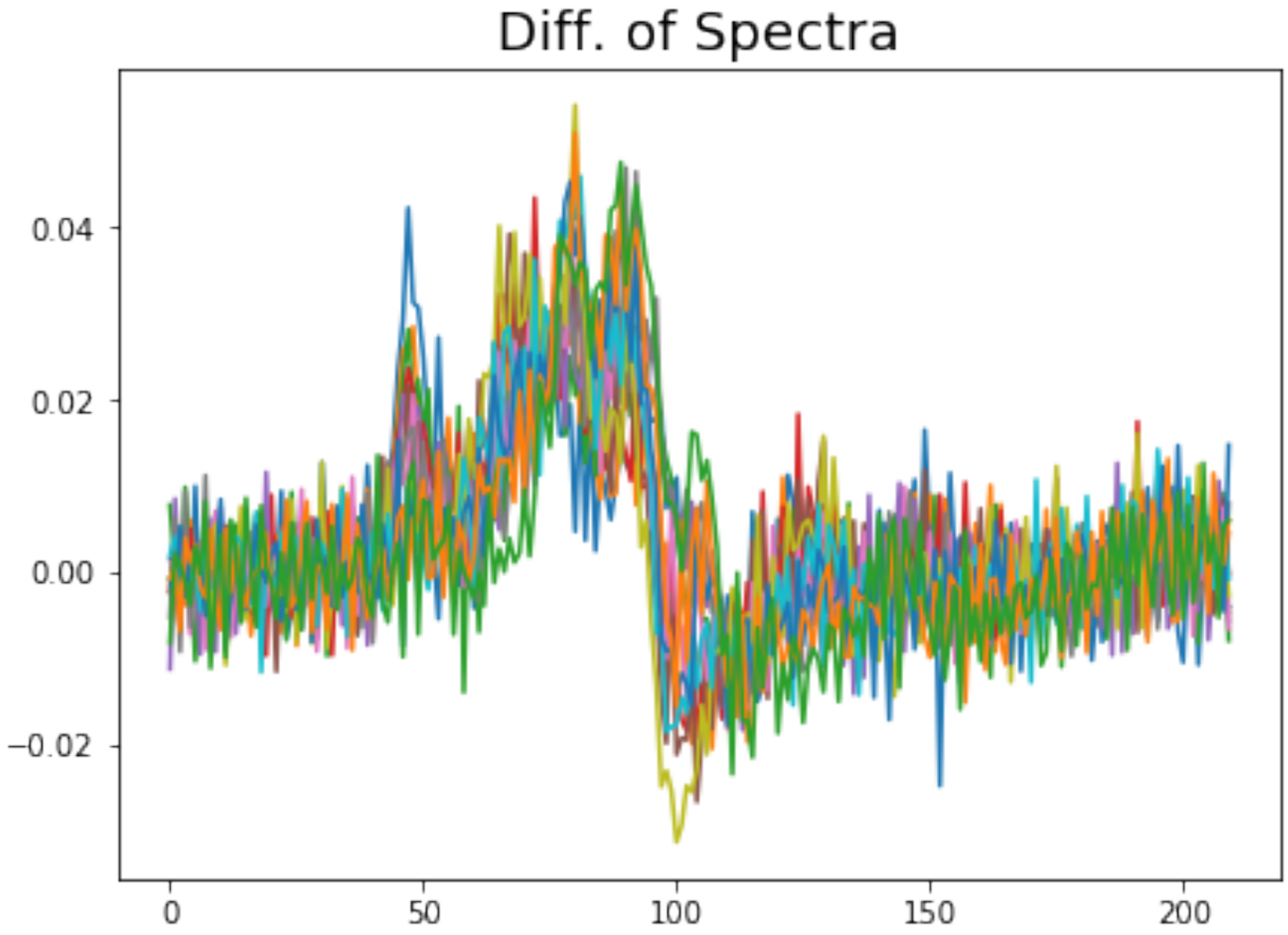
Fe, FeO, Fe_2O_3 の混在する領域を認識させるため

ノイズパターンを与える



- 様々なノイズパターンを乱数で与える(1000パターン/クラス)
  → ノイズに強くなる
- S/N は測定データにレベルを合わせている
  S/N にバリエーションを与えると、認識度が低下する

ノイズパターンの微分



学習データ ＆ テストデータを用意

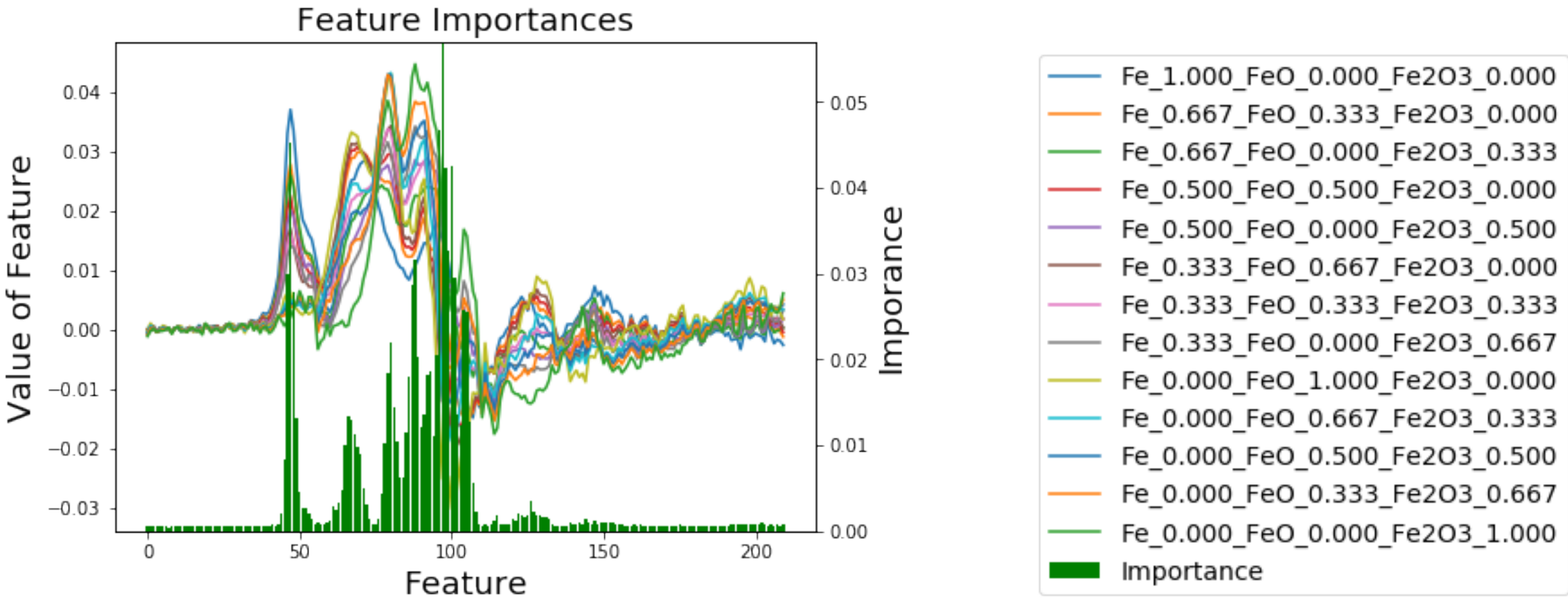
ランダムフォレストの
パラメータ

RandomForestClassifier(n_estimators=100)

n_estimators=100　決定木の数 (多ければ多いほど良い)

- 多いほどマシンパワーを消費する
- いずれ性能が頭打ちになる

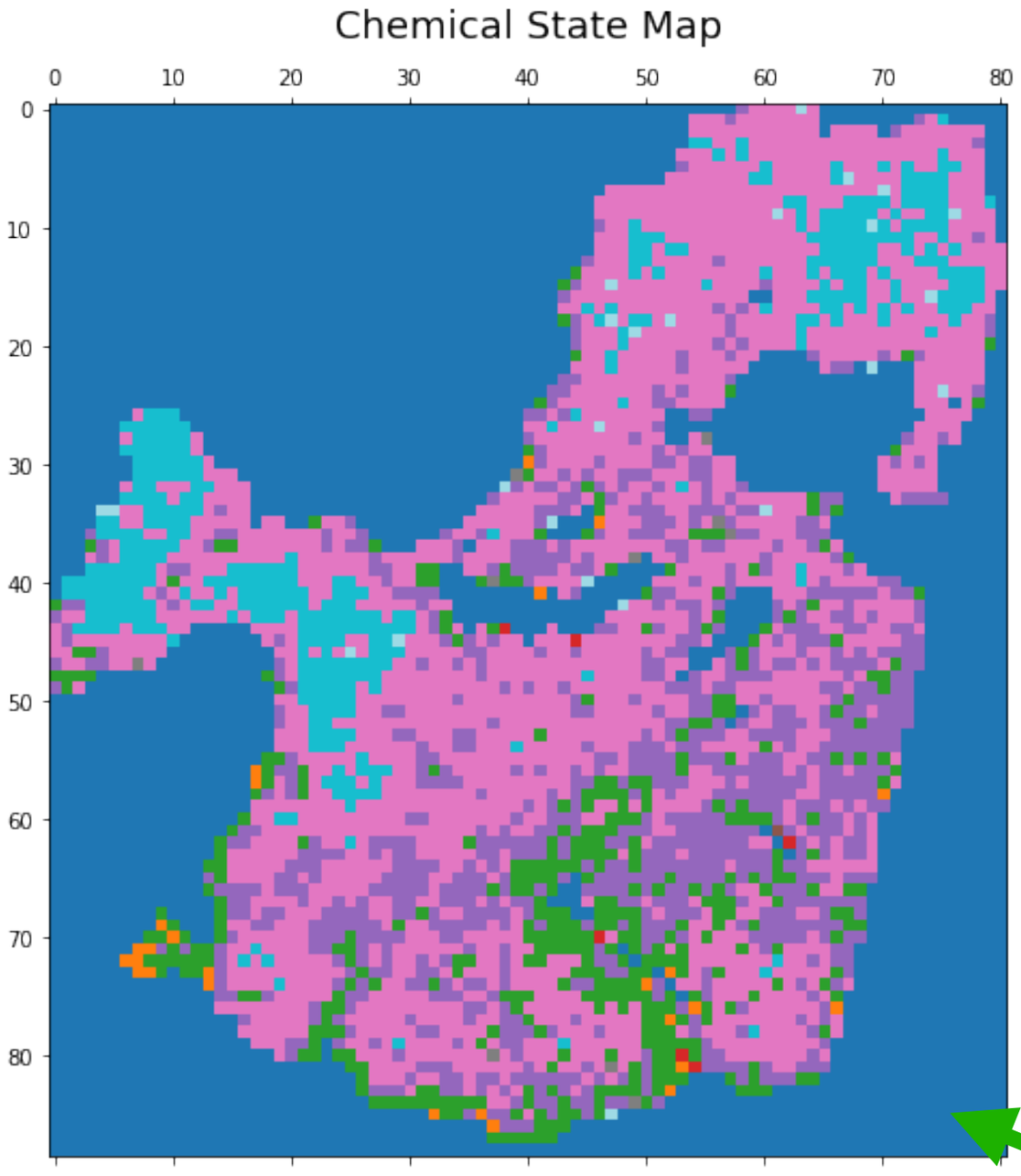
特徴量の重要度



変化の乏しい特徴量は注目されない

分類に使用されない

Fe化学状態マップ



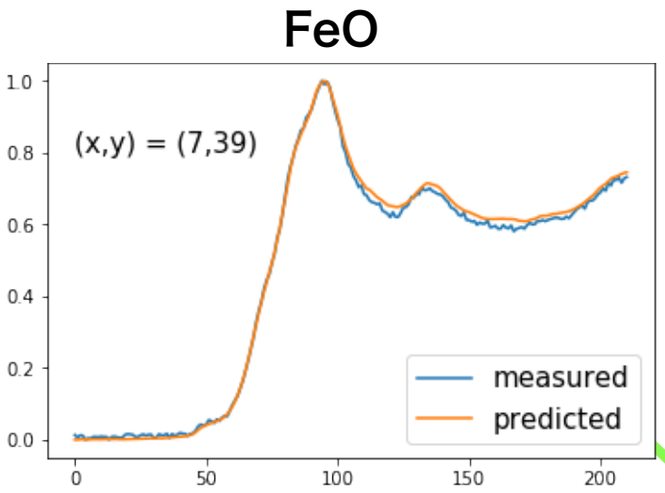
混在比

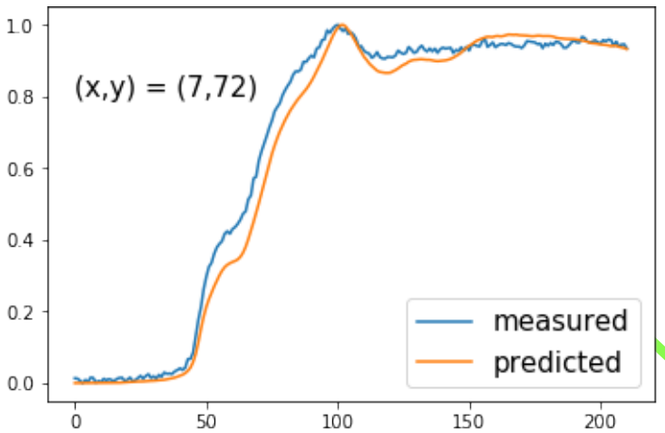
	Fe	FeO	Fe_2O_3
	0	2	1
	0	1	0
	1	0	2
	1	1	1
	1	2	0
	1	0	1
	1	1	0
	2	0	1
	2	1	0
	1	0	0
	BG		

BG領域としてマスク

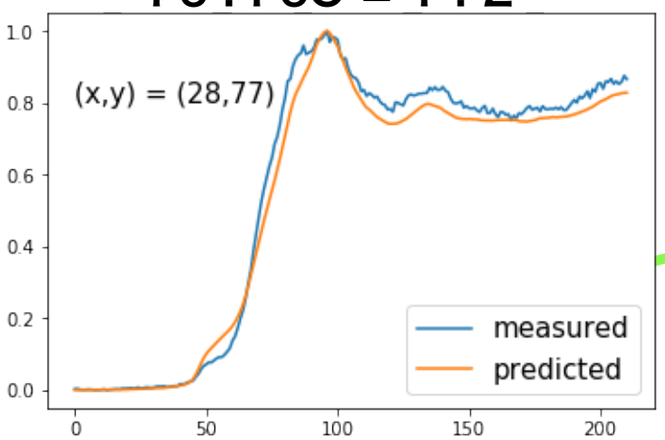
(*) 解析前に隣接 10 × 10 pix を平均化

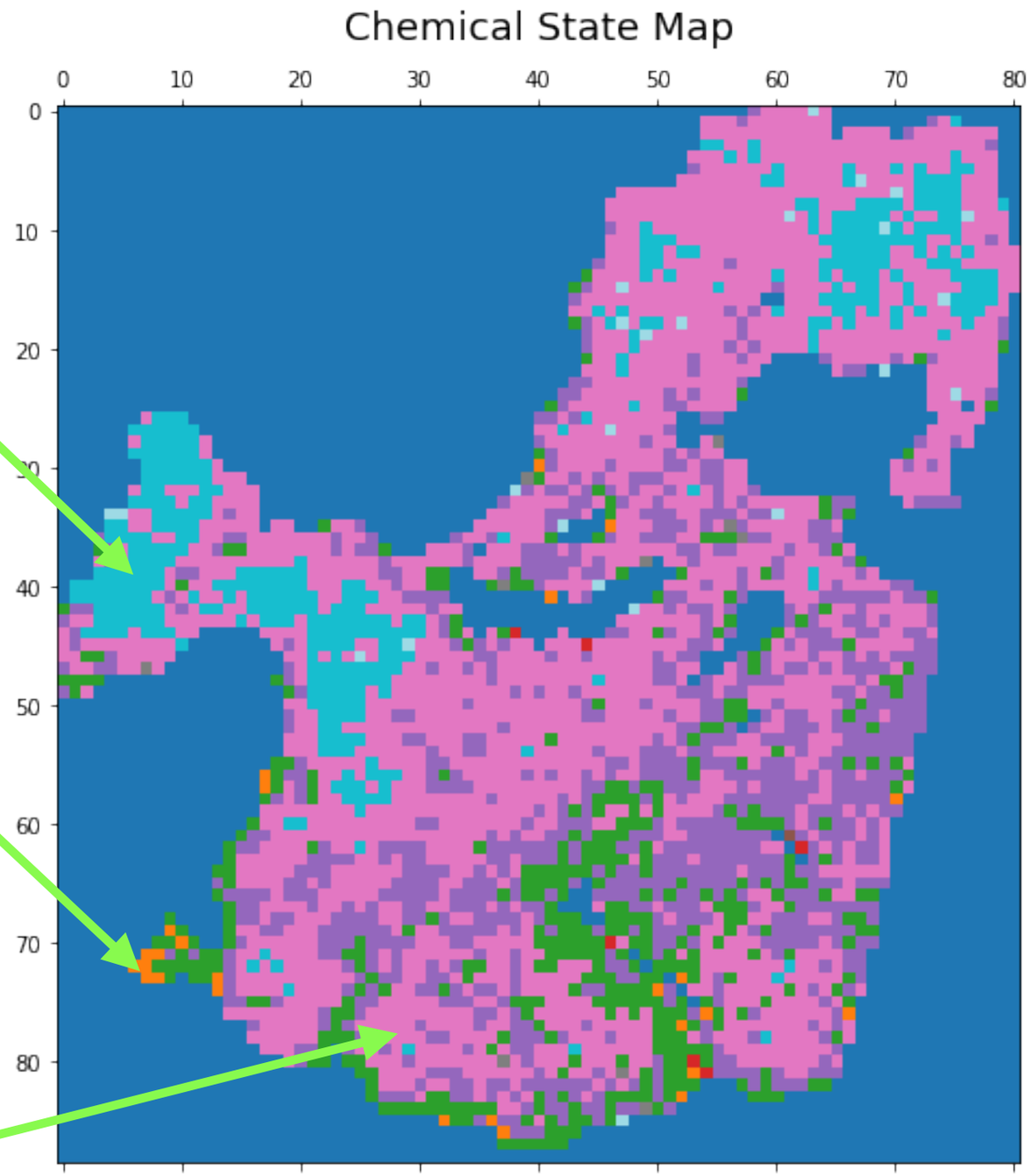
測定データとクラスの比較





Fe : FeO = 1 : 2

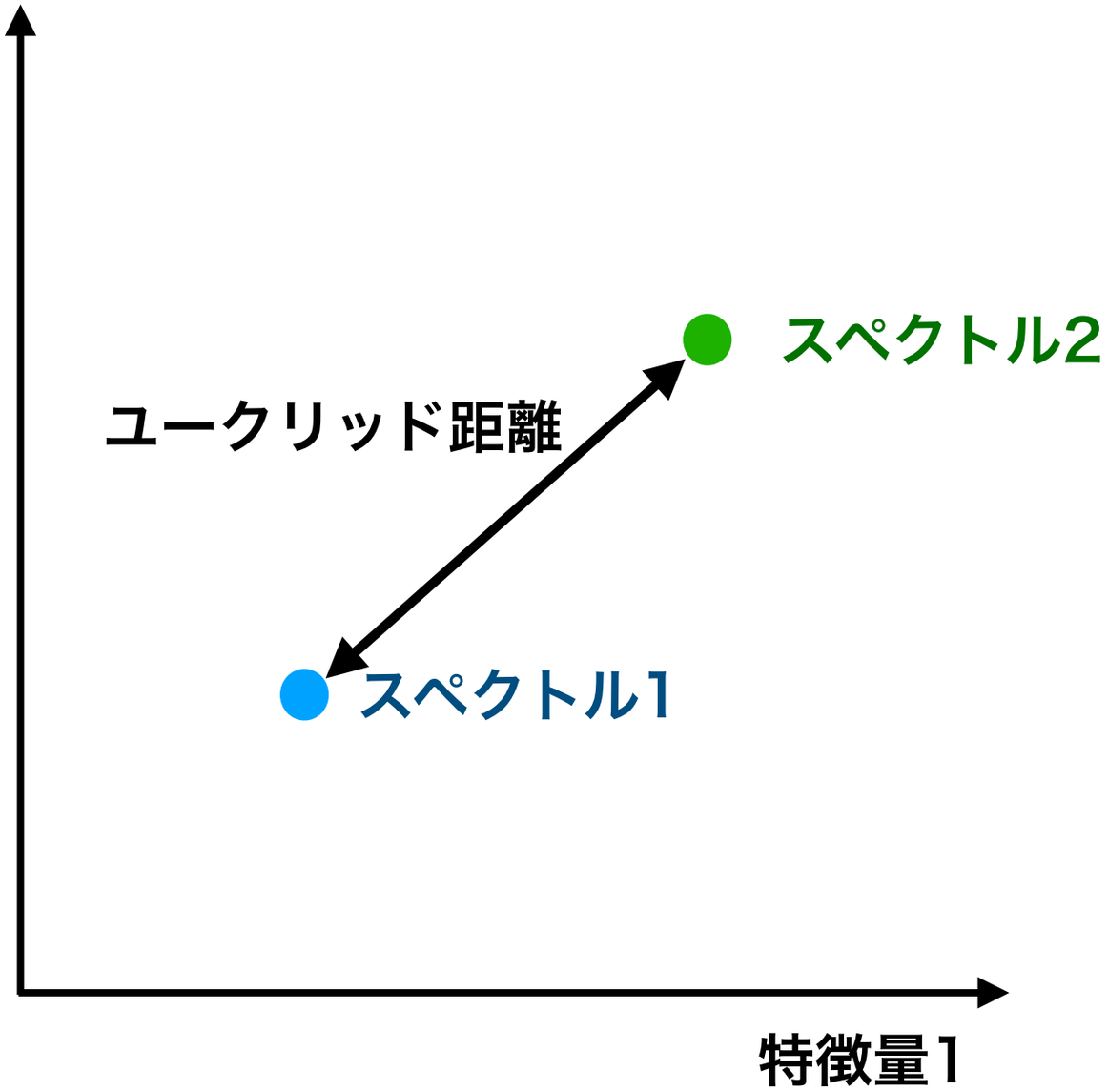




混在比

Fe	FeO	Fe_2O_3
0	2	1
0	1	0
1	0	2
1	1	1
1	2	0
1	0	1
1	1	0
2	0	1
2	1	0
1	0	0
BG		

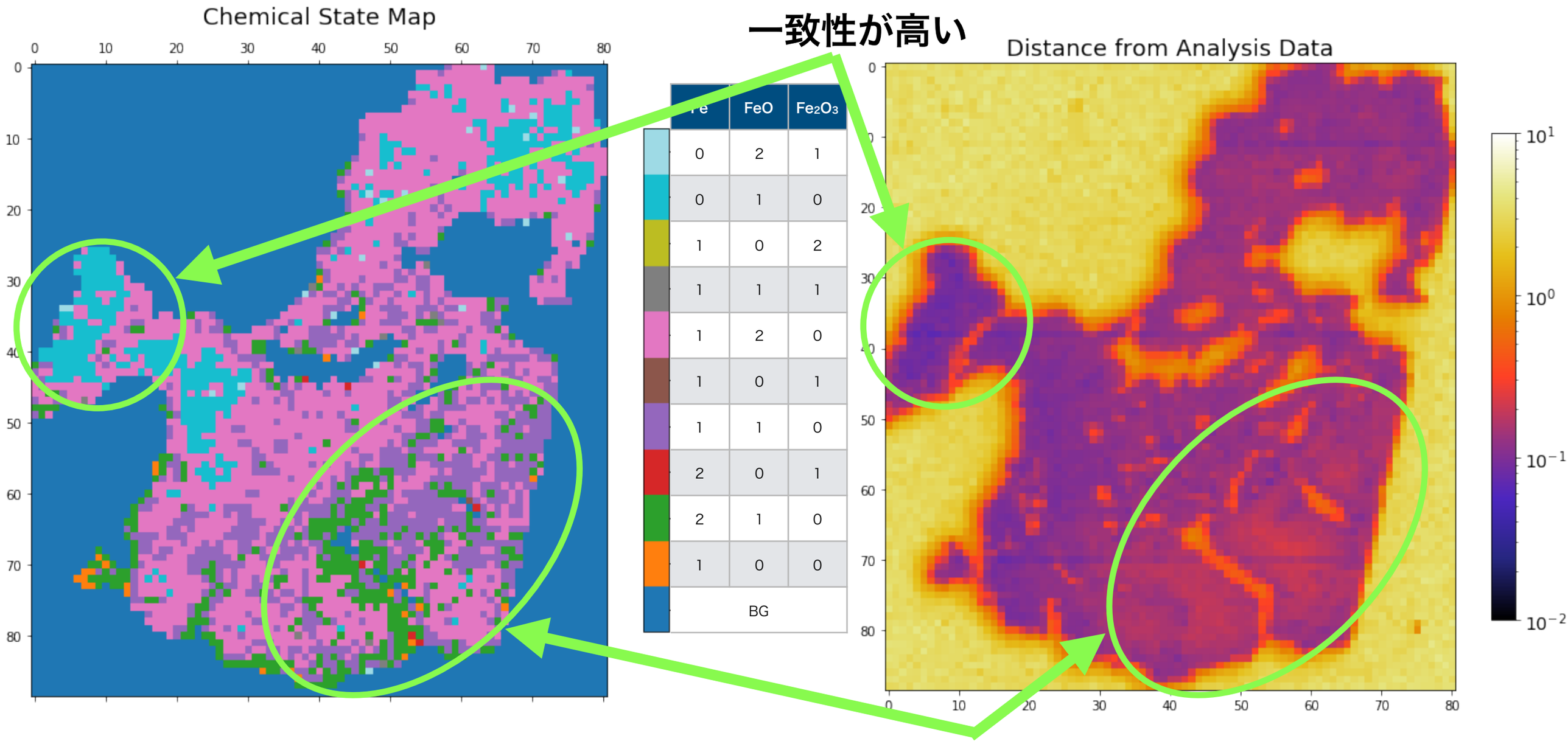
特徴量空間における距離



特徴量が2つの場合

- 210のエネルギー点→210次元の特徴量空間
- 1つのスペクトルは特徴量空間の1点で表される
- 距離が近いほど「似た」スペクトル
- ユークリッド距離以外の距離も考えられる

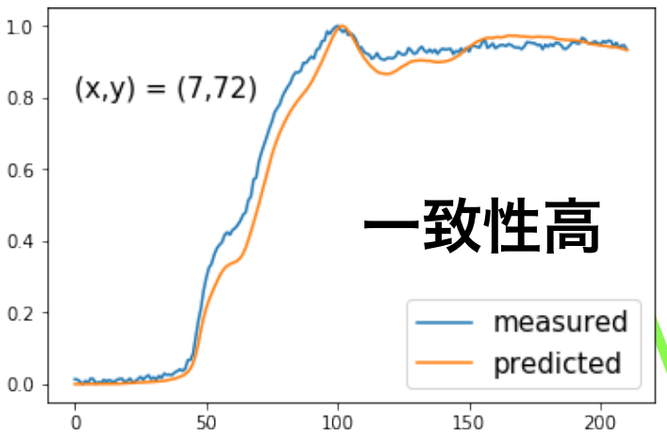
測定データとクラスの距離



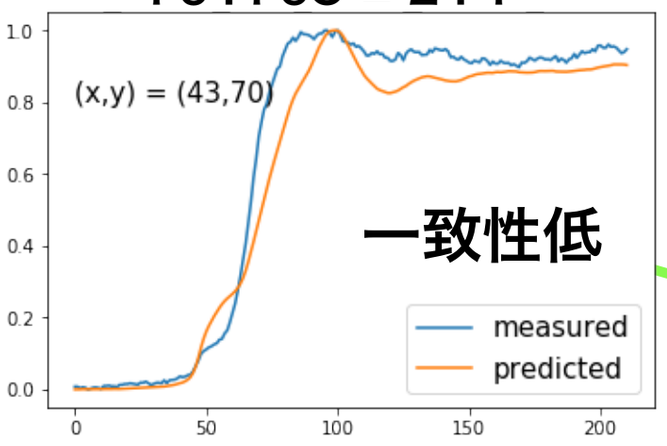
“Fe のプロファイルが強い” と推定した領域は、
取り扱いに注意を要する

定義したクラスの限界

Fe

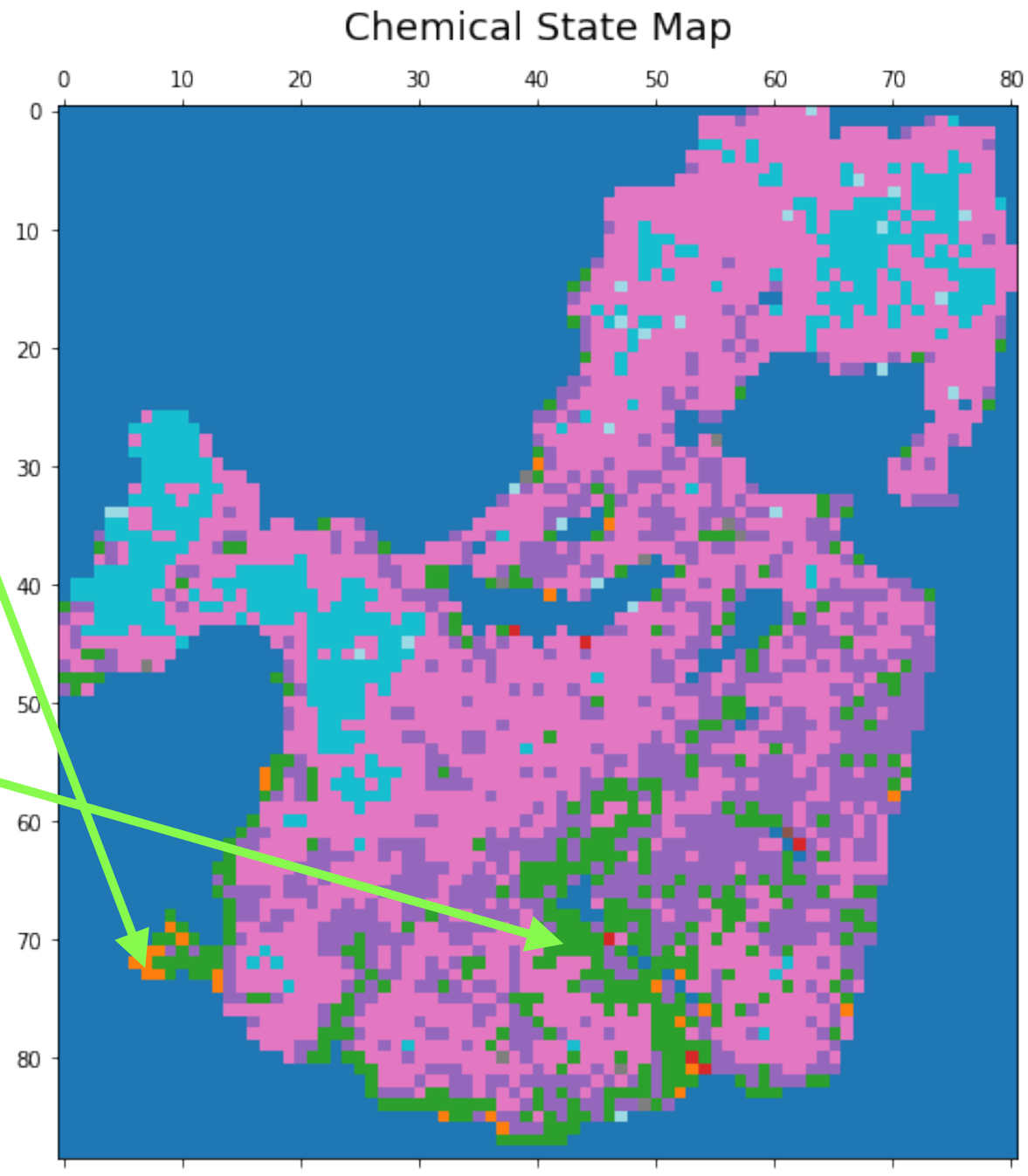


Fe : FeO = 2 : 1



未知の要素の存在

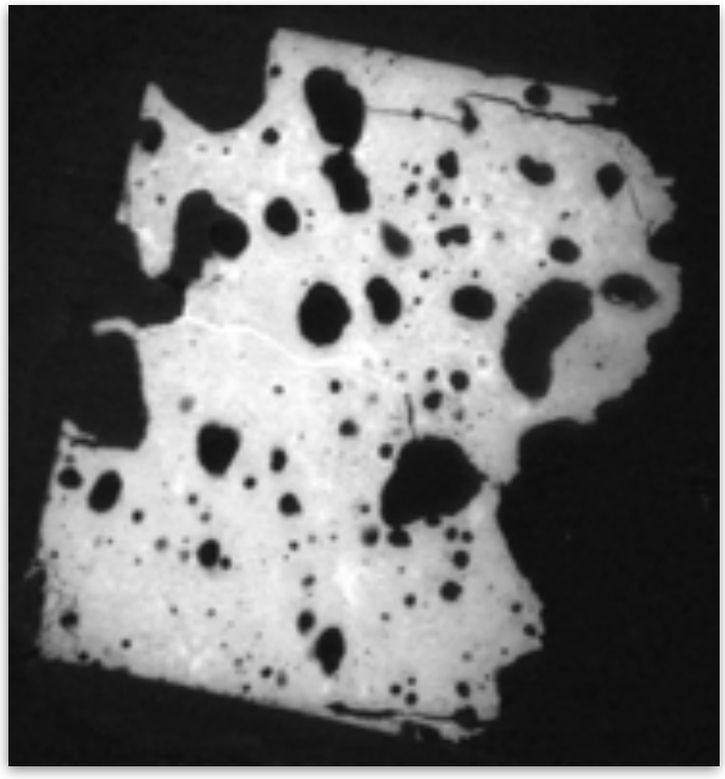
が示唆される



混在比

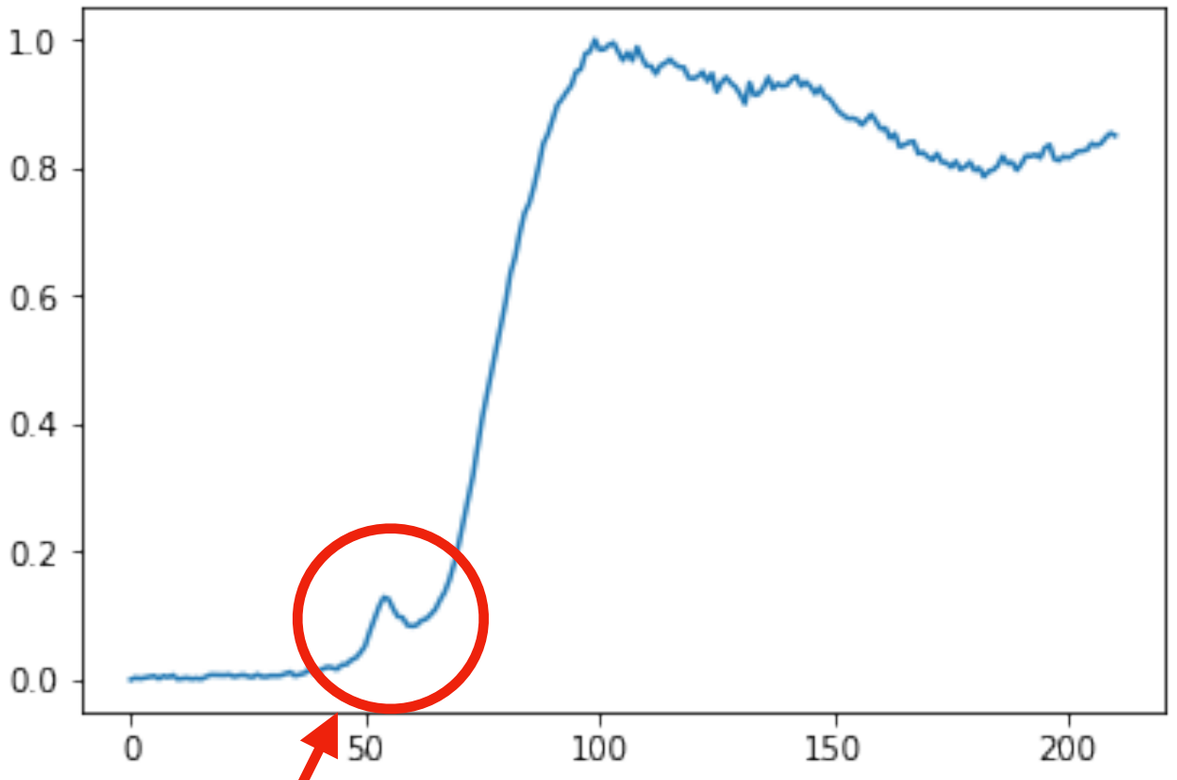
Fe	FeO	Fe_2O_3
0	2	1
0	1	0
1	0	2
1	1	1
1	2	0
1	0	1
1	1	0
2	0	1
2	1	0
1	0	0
BG		

試料番号:0



還元処理なし

あるピクセルのスペクトル



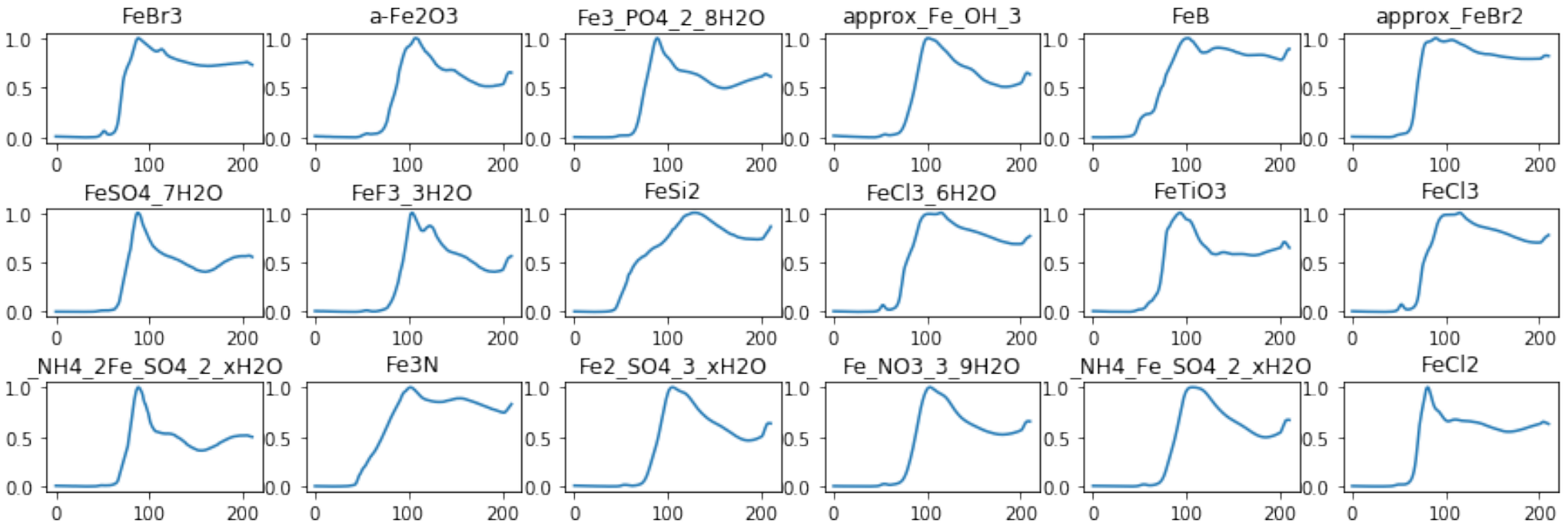
プリエッジピーク

Fe, FeO, Fe_2O_3 では説明できない

(未知の主要要素がある)

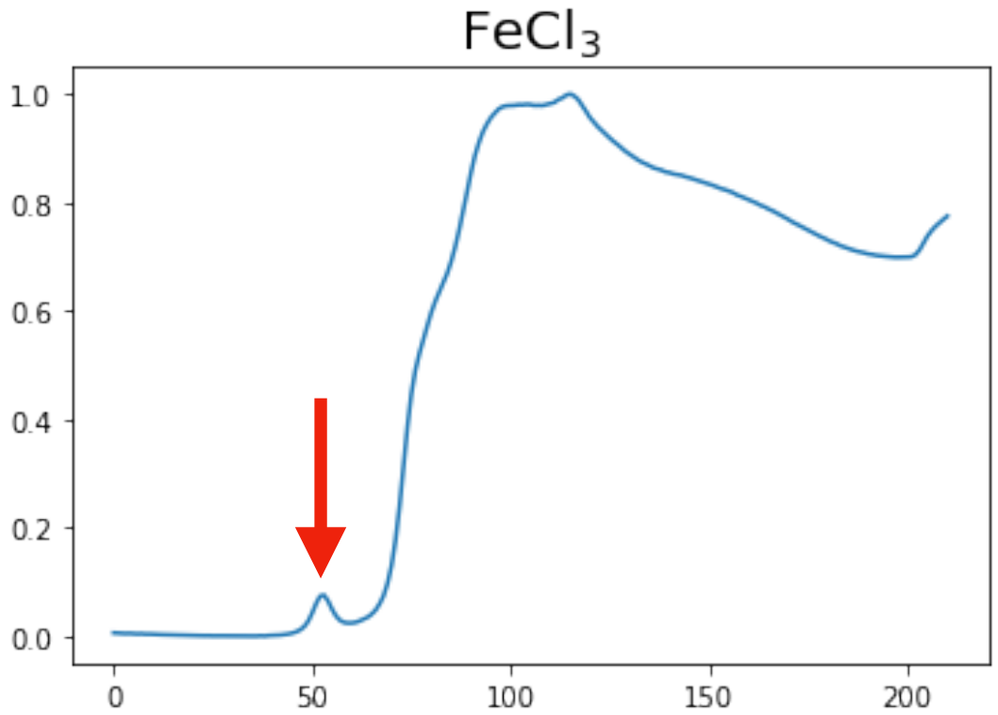
BL14B2 標準試料データベース

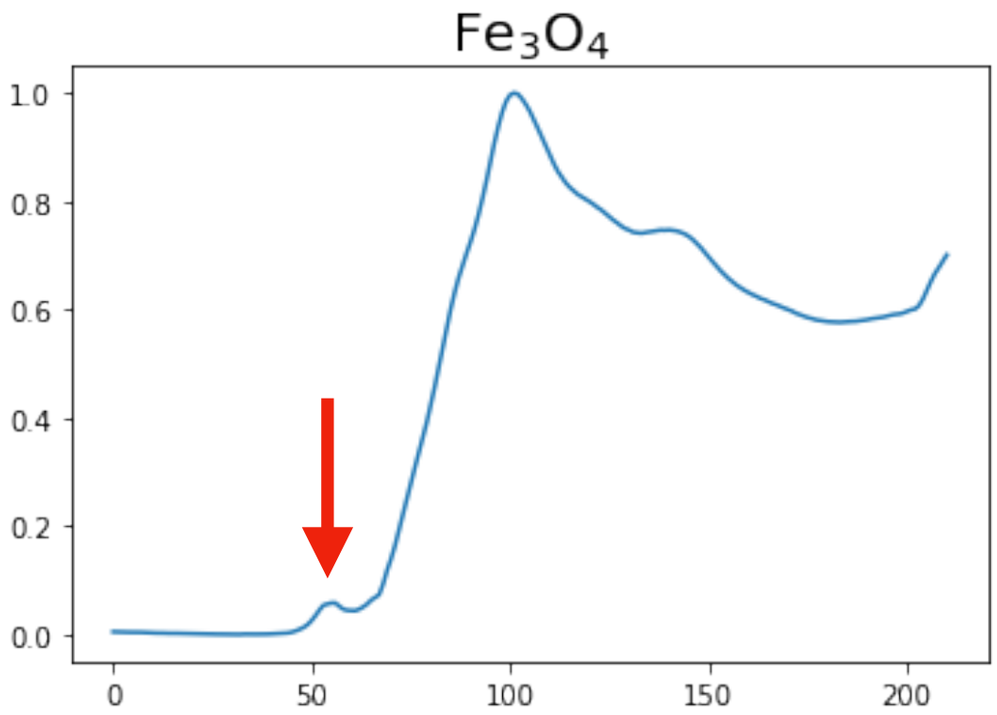
https://benten.spring8.or.jp



Fe-K 端スペクトルを全てダウンロード

プリエッジピークを持つ
スペクトルをピックアップ





これを元に専門家の意見を聞く

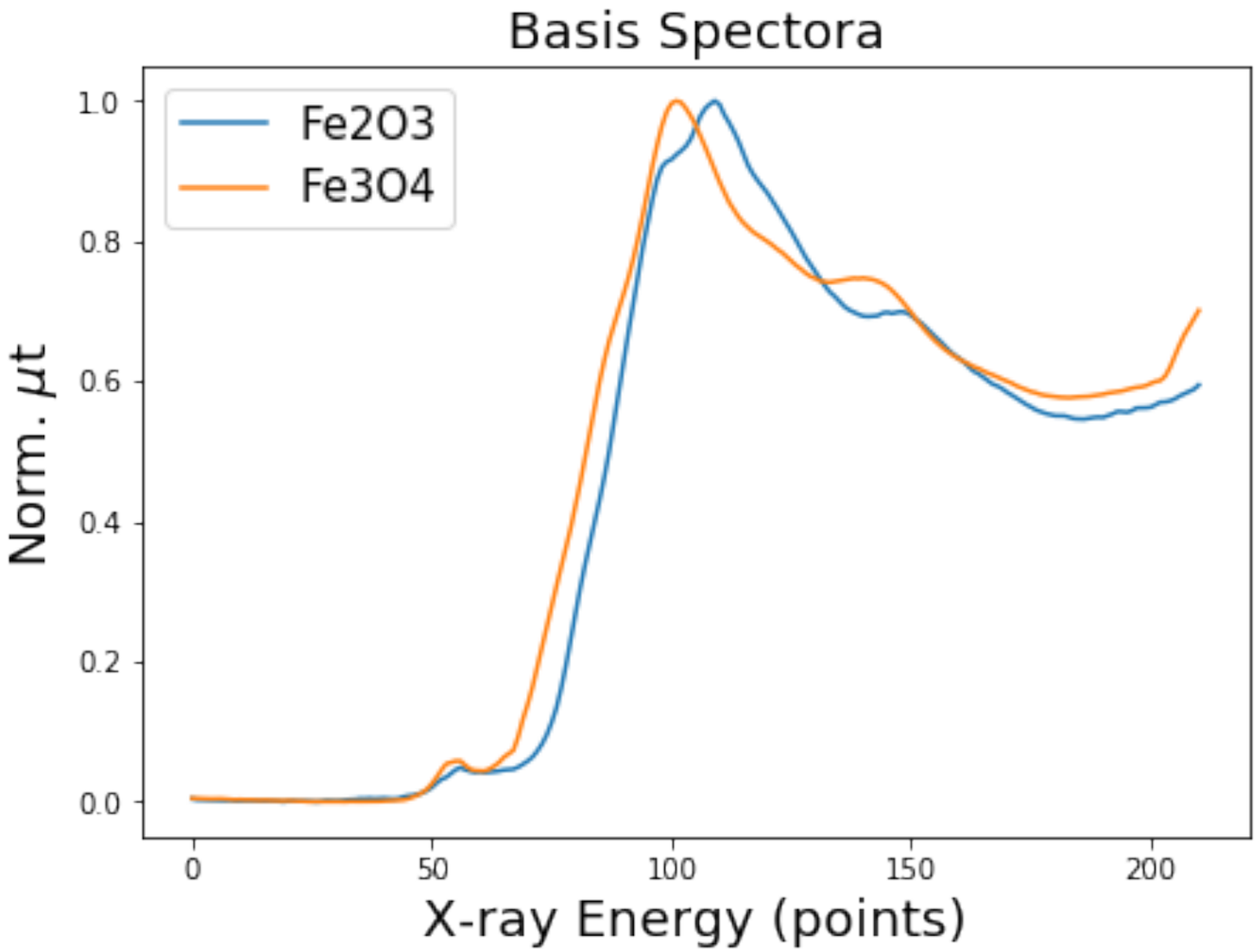
新しい先験的知識

- 試料提供者「×$FeCl_3$、◎Fe_3O_4」

- 他の専門家「FeOは(ほとんど)存在しないと思われる」

- 還元処理をしていないので、Feはないと考えられる

新しい基底スペクトル

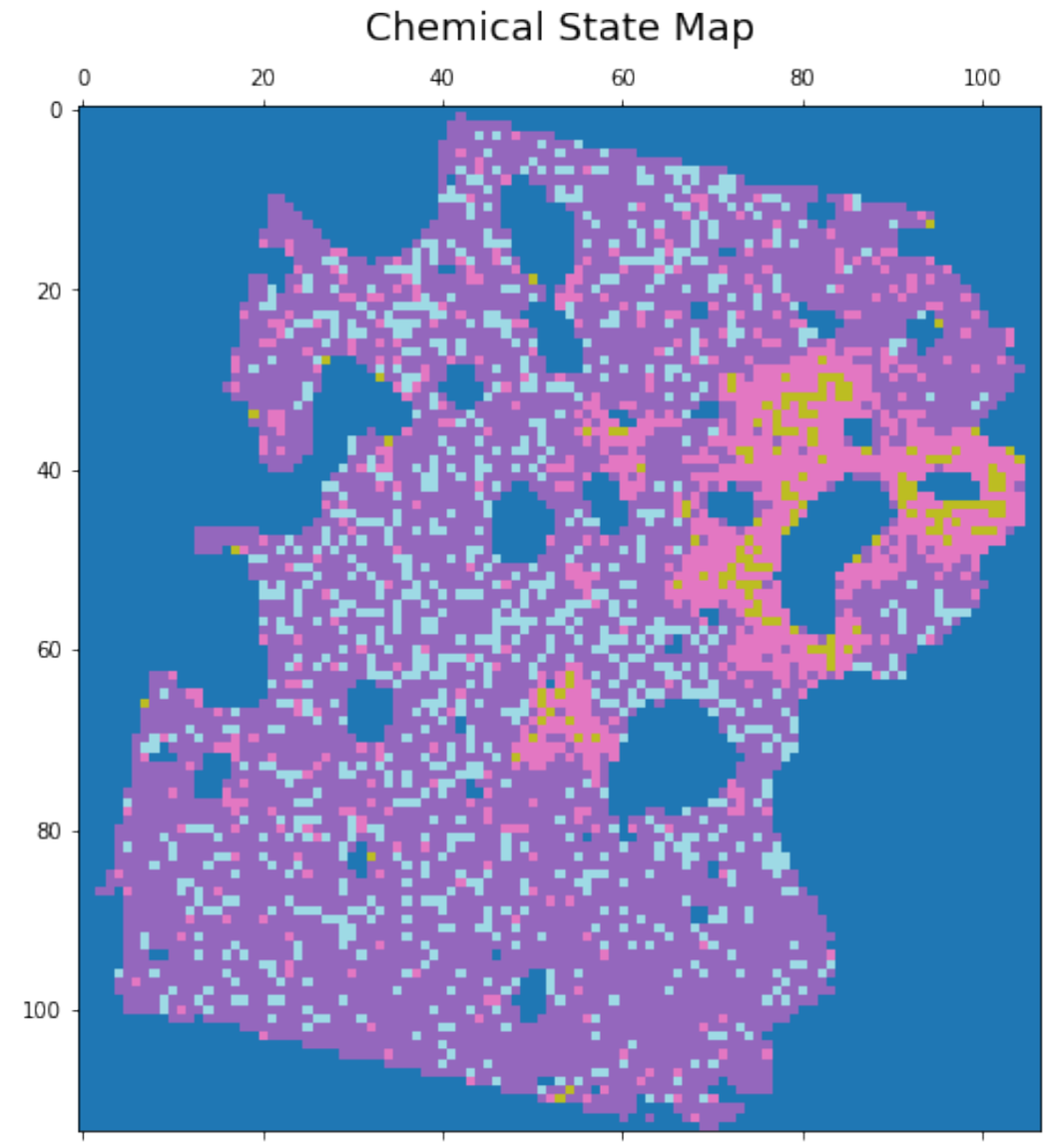


ヘマタイト　マグネタイト

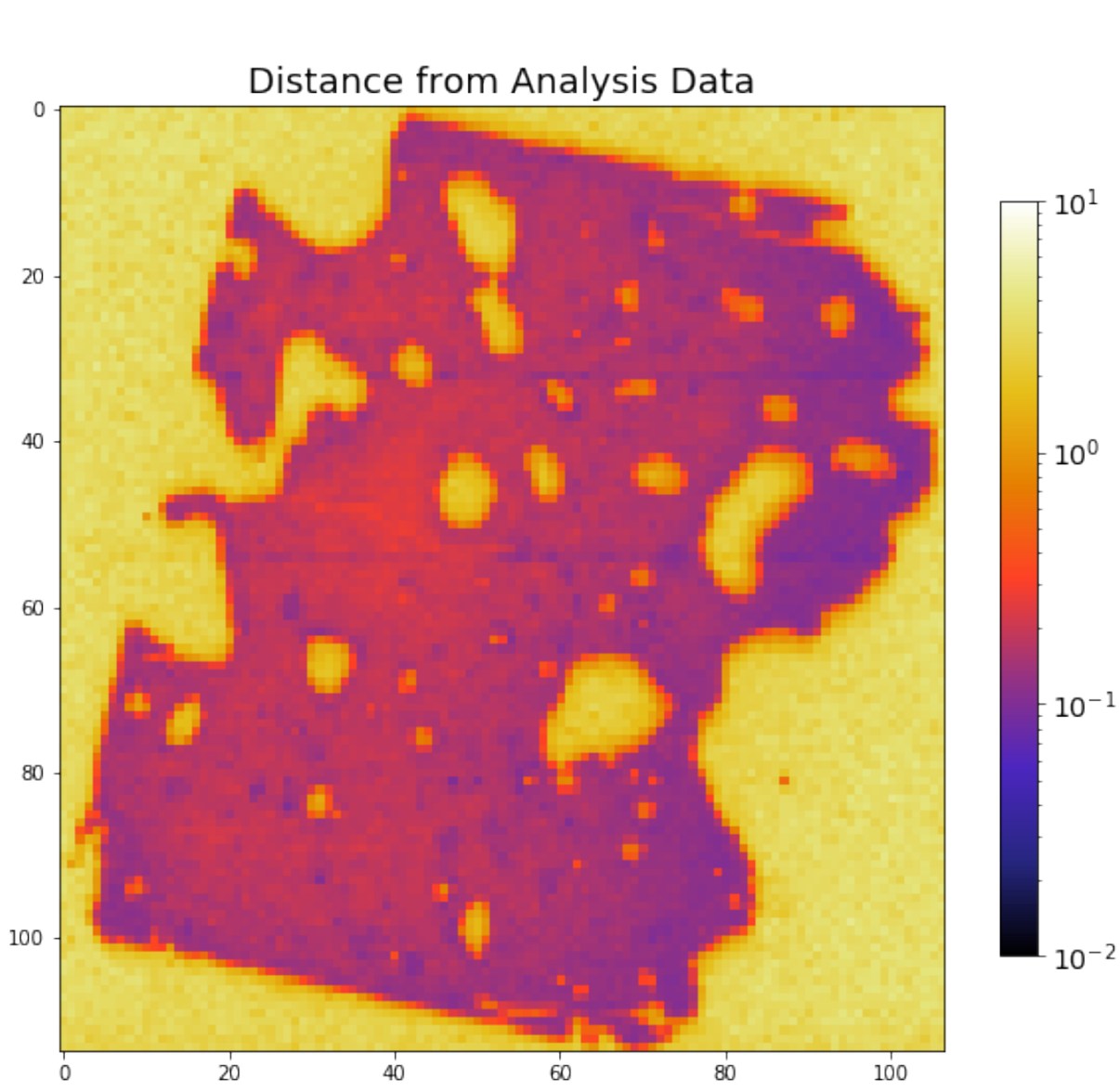
- Fe_2O_3, Fe_3O_4 のみで構成
- Fe_3O_4 はDBのを使用
- ↑Fe_2O_3 に合わせてエネルギー校正/エネルギー範囲の切り出し

(注意)
・DBと手持ちのデータで共通のスペクトルがないと校正不可
・測定条件が共通化されているからエネルギー範囲を合わせられる

測定データとクラスの距離



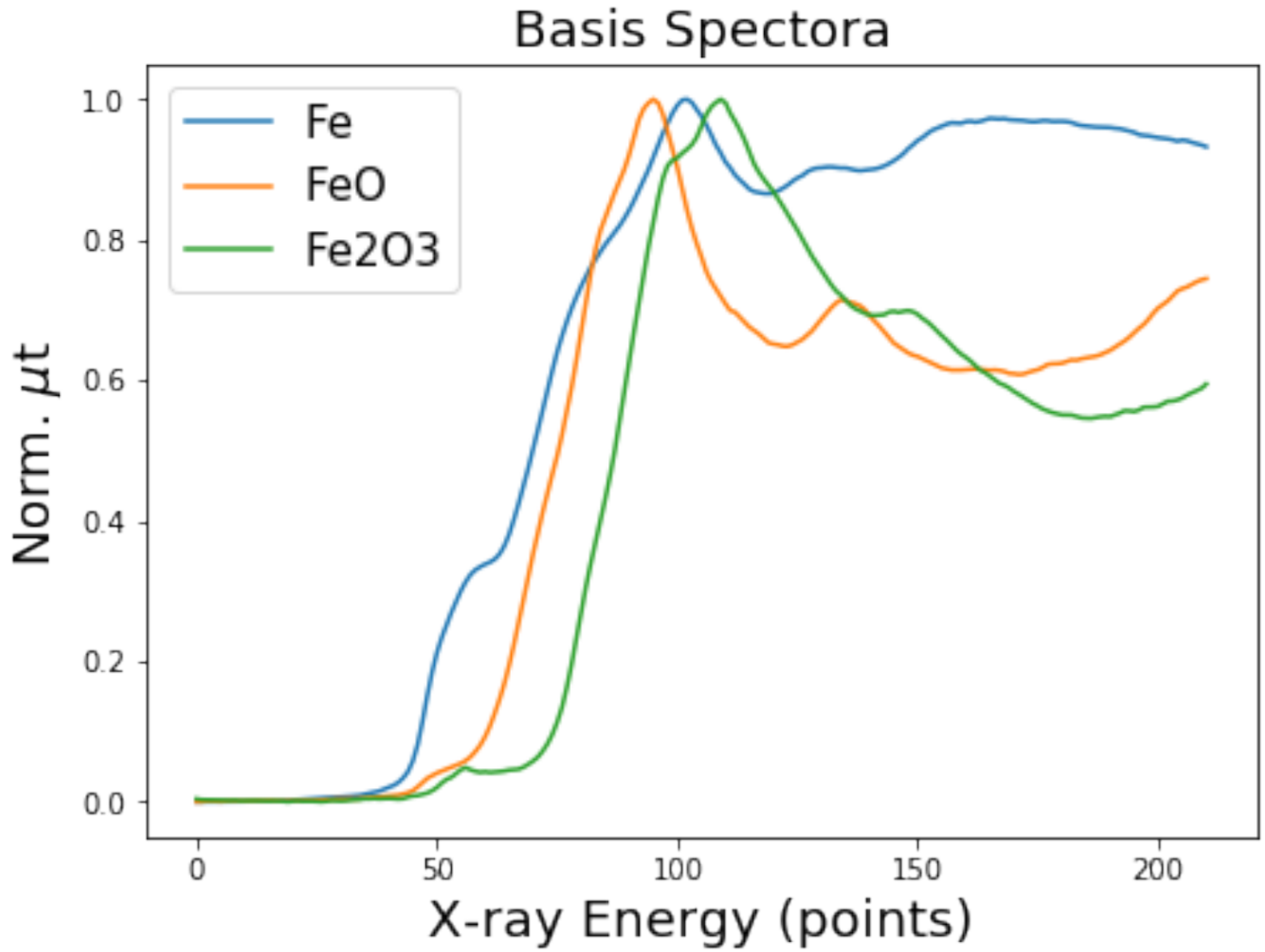
	Fe_2O_3	Fe_3O_4
	0	1
	3	1
	1	1
	1	3
	1	0
BG		



- 試料番号6に比べて一致度はやや低い(**未知の主要要素がまだある**)
- Fe_2O_3が濃い領域は比較的一致度が高い

危険なケース

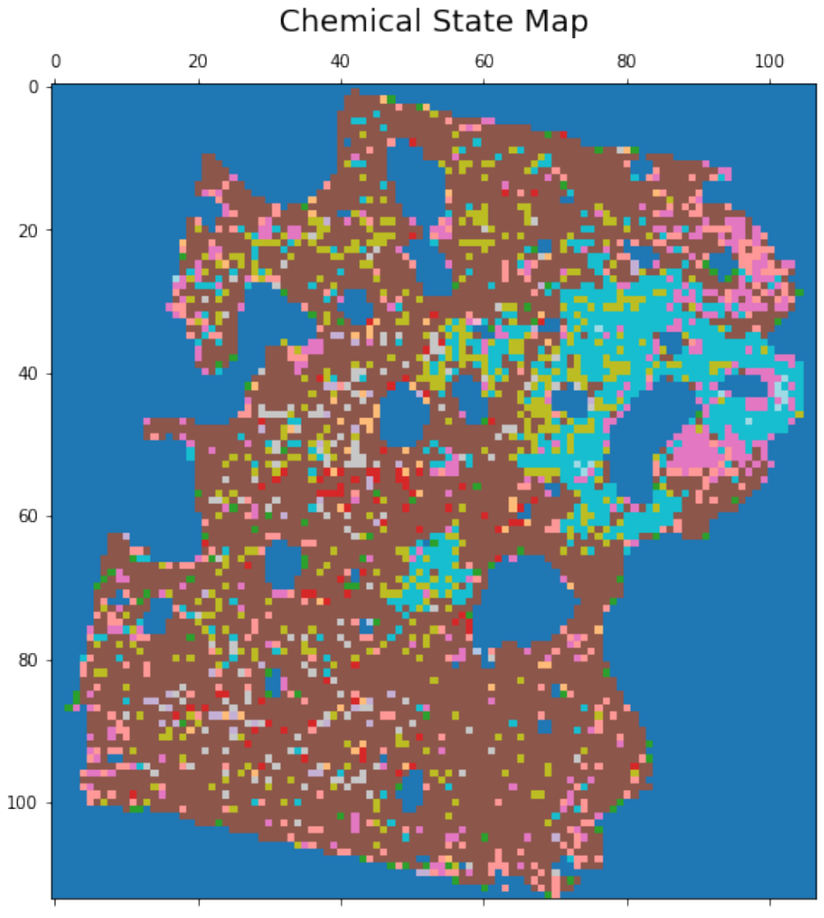
不合理なクラス定義による危険な事例



還元処理していない試料(0番)に

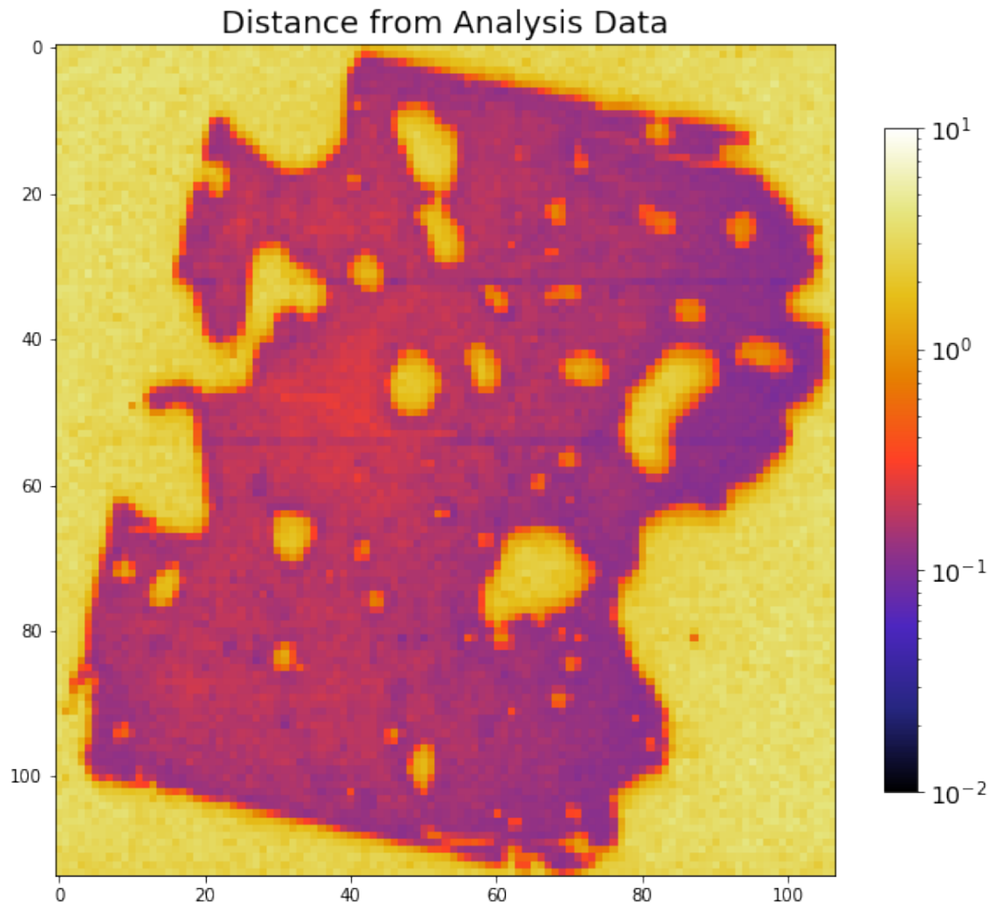
高還元率試料(6番)と同じ基底スペクトルを使ったら?

測定データとクラスの距離



混在比

	Fe	FeO	Fe_2O_3
	0	0	1
	0	1	2
	0	1	1
	0	2	1
	0	1	0
	1	0	2
	1	1	1
	1	2	0
	1	0	1
	1	1	0
	2	0	1
	2	1	0
	1	0	0
	BG		



合理的クラス定義と一致度がほとんど変わらない

多くの基底を入れればそれだけ一致しやすくなる

今後の課題

- ユークリッド距離以外の距離

  マンハッタン距離、Jensen-Shanon 情報量、 …

- 距離以外の評価法の模索、多角的評価

- 効果のある/ないケースの見極め

⋮

効果的な解析/評価法の組み合わせ

新たに湧き上がった疑問

「より少ないパラメータで説明できる方が尤もらしい」

というメタ的判断

我々は、パラメータ数以外の要素も合わせて

多角的に判断している。

種類も数も流動的

「モデルの合理性」を機械的に評価できるか?

「モデルの合理性評価」のモデル化?

まとめ

- RandomForest によるXAFS指紋法の実装
  - Fe化学状態の可視化

- 合理的なクラス定義には先験的知識が重要
  - データをよく眺める
  - より多くの情報源を持つ (専門家の意見, DB, …)
  - 複数の解析結果からのフィードバック

- 解析結果の評価法はさらなる検討が必要