
「データ構造化ワークショップ」へようこそ

2024年12月4日 ARIMアカデミー データ構造化ワークショップ2024

物質・材料研究機構
マテリアル先端リサーチインフラセンターハブ

松波 成行

簡単な自己紹介

1998年 北海道大学大学院地球環境科学研究科博士課程修了。博士(地球環境科学)。

新規導電性高分子の合成とNMRによる分子運動・物性研究

1998年～2015年 民間会社

Ziegler-Natta触媒開発と工業化
有機ELのディスプレイデバイス開発

2015年 物質・材料研究機構 調査分析室 室長

2017年 同 統合型材料開発・情報基盤部門 参事役

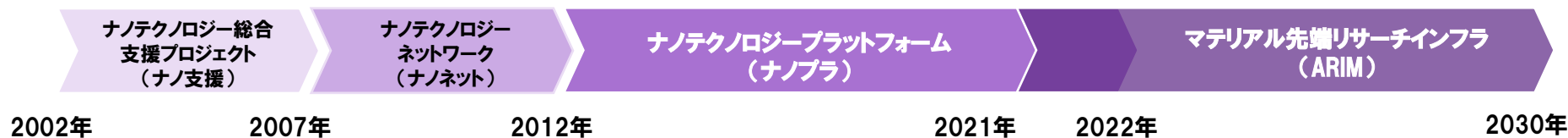
2022年 現職



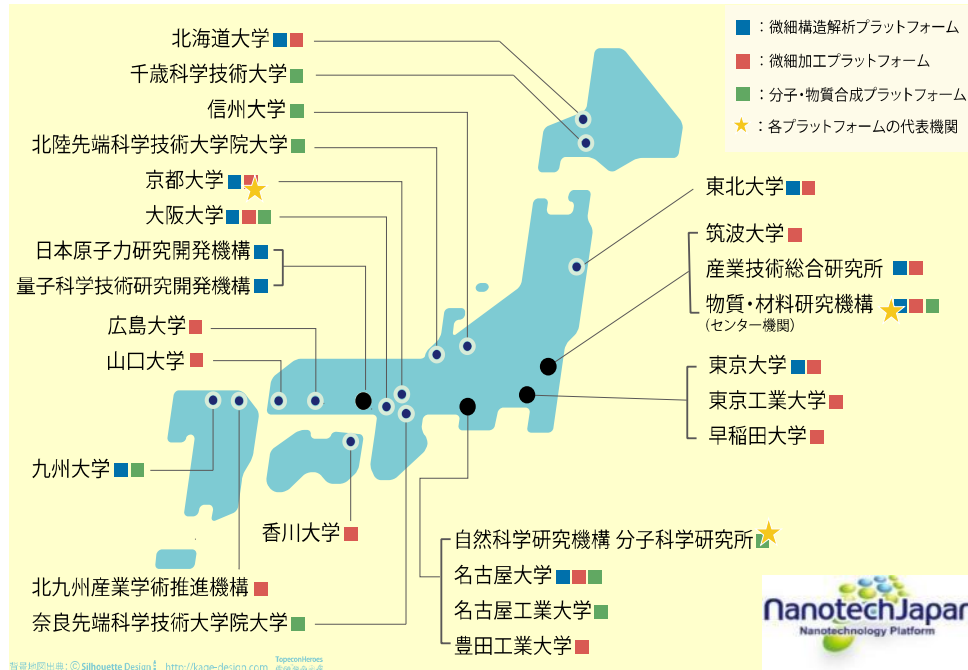
- | | |
|--------------|-----------------------------------|
| ・ データインフラ | : 実験系データ基盤(DX)のネットワーク・システム設計 |
| ・ データアーキテクチャ | : データ収集・蓄積・共用にかかるデータ構造化設計 |
| ・ データルール | : 実験系データにかかるデータマネジメント・データ規程整備 |
| ・ ARIM事業 | : 全国25機関の設備共用装置(約700台)からのデータ運用の統括 |
| ・ プログラムスキル | : python歴は5年ほど (Rは2年ほど) |

1. マテリアル先端リサーチインフ事業について

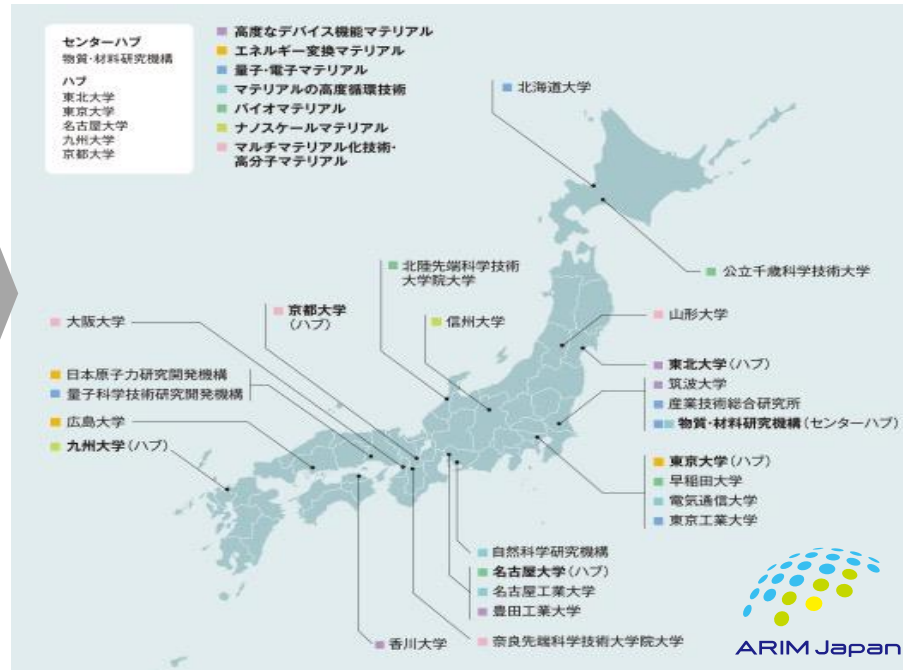
MEXT共用支援事業におけるARIM



ナノテクノロジープラットフォーム (3PF : 25法人、37実施機関+センター機関)



マテリアル先端リサーチインフラ (7重要技術領域 : 25法人 : 1センターハブ、5ハブ、19スポーク)



設備共用

設備共用
データ共用

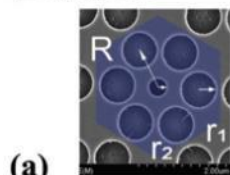
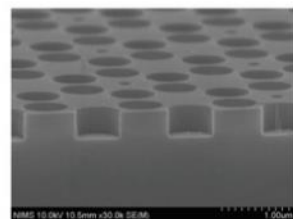
ARIMの多彩な共用機器

微細加工系装置

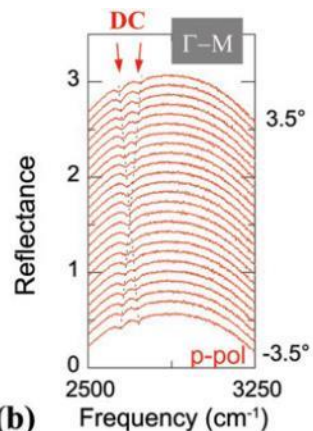
原子層堆積(ALD)装置



電子ビーム描画装置



(a)



(b)

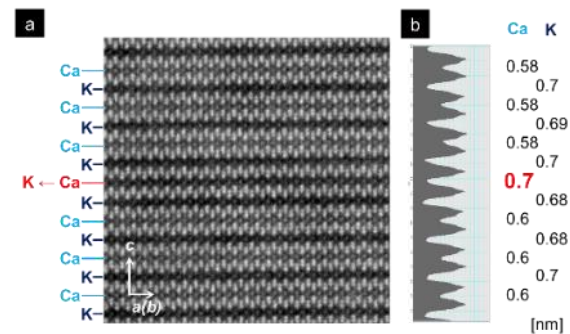
SOIフォトニック結晶

先端計測系装置

300kV収差補正電子顕微鏡

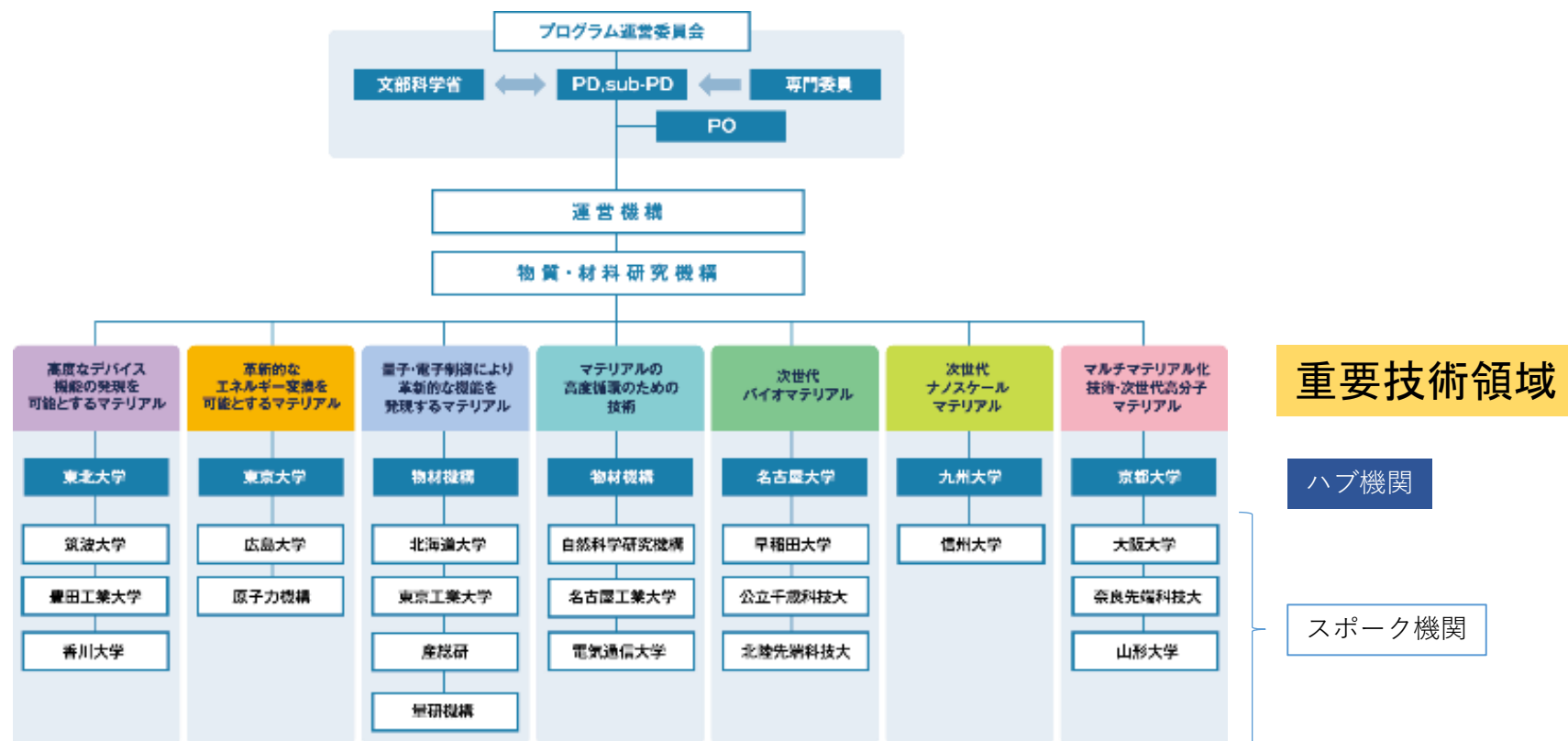


ラマン顕微鏡



鉄系超電導材料のHAADF-STEM像及び層間距離の計測

ARIMの体制

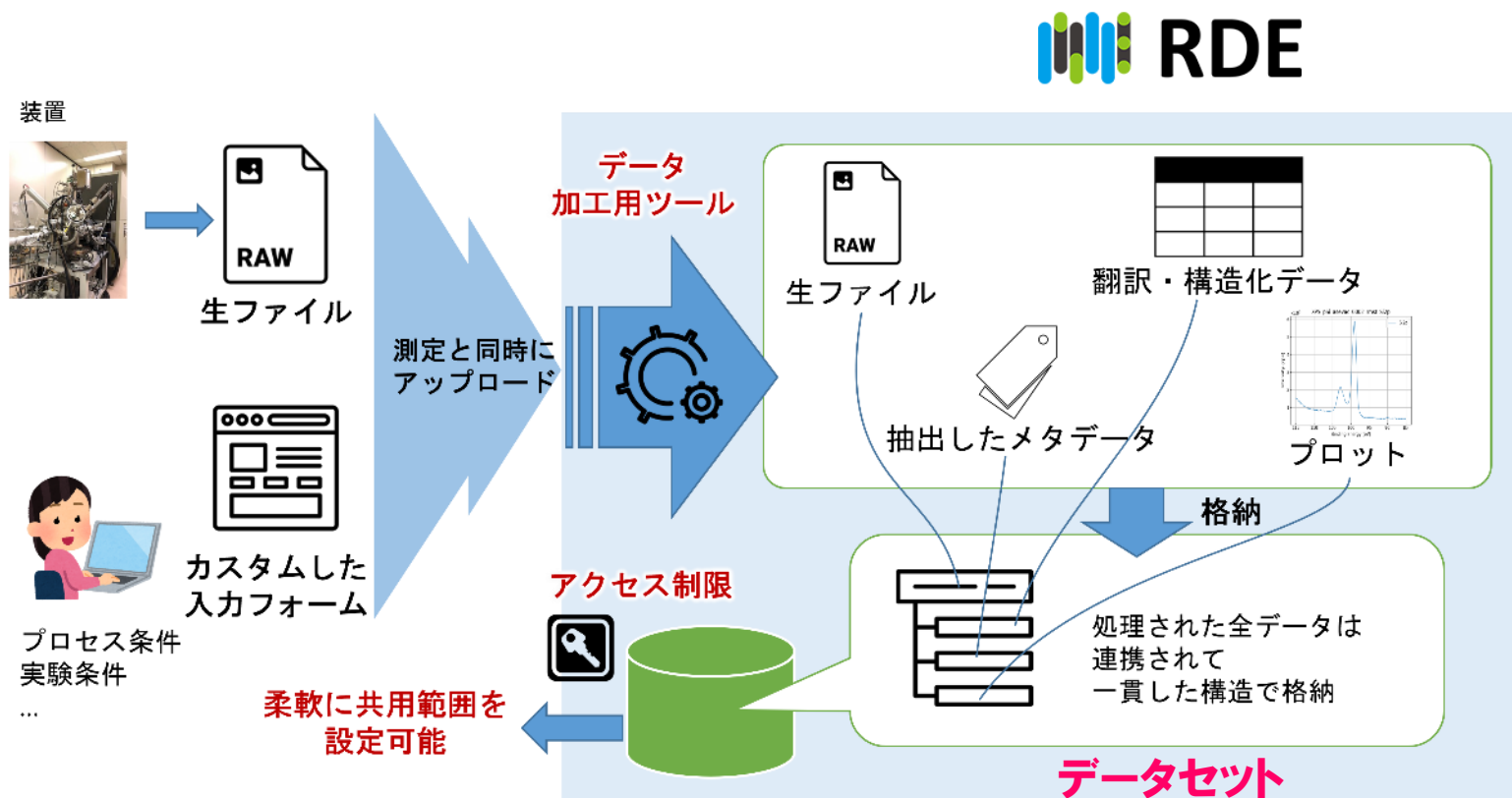


DX基盤でマテリアルデータを蓄積し、データ駆動型研究の支援します

2. データ登録（データ構造化）

データ構造化システム(RDE)の概要

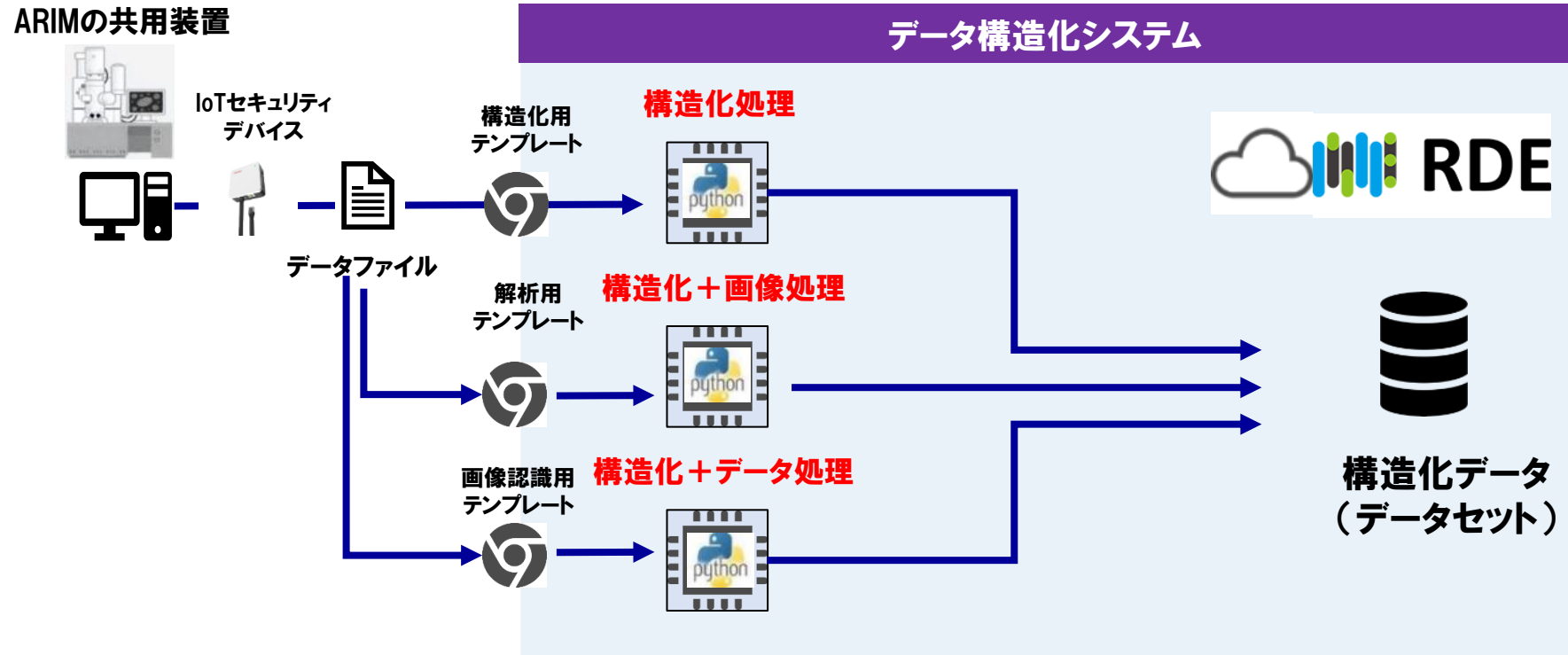
- ARIMでは共用機器等からのデータをワンストップで**データ構造化**を行い、機械学習やデータ駆動型研究へ利活用しやすい「**データセット**」としてデータ利用者様にご提供します。



RDEはNIMSが独自に開発したデータ構造化システムで、システム用語では「**データウェアハウス(DWH)**」の位置づけになります。

ARIMのDX戦略： AI Readyのデータセット提供

- ARIM共用機器等からの出力データのデータ構造化ツールを整備
- アップロードするだけでAIに使いやすいデータセットとして自動で構造化。

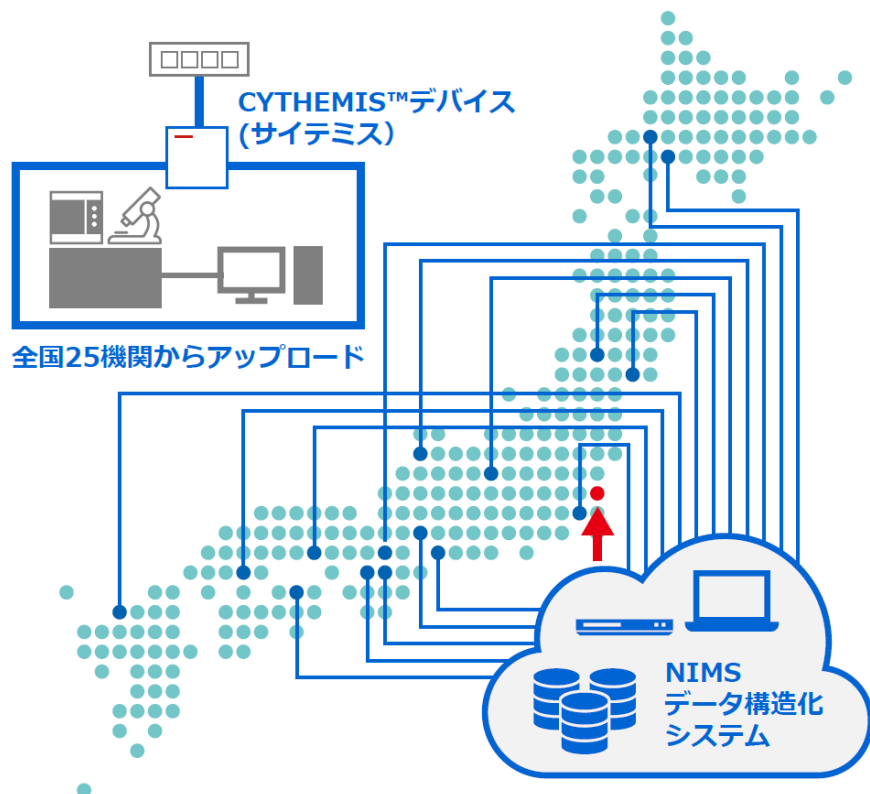


インフォーマティクに必要な構造化パイプラインを整備
→ みなさんのデータ駆動型研究の支援

日本全国からのデータ登録体系の確立

- 2023年度よりARIM機関からセキュアな環境のもとデータ登録が促進される基盤運用をスタート

2024.10月末



登録ユーザー数

約3,500人

データセット
開設数

約8,000

ファイル総数

約660,000

総ファイルサイズ

約1,100GB

データ構造化自動対応が進んだことによりデータ登録が飛躍的に加速

3. データ共有

ARIMデータポータルへのデータ共有

ARIM-RDE (データウェアハウス)

構造化データ
(データセット)



- ① rawデータ (非公開/非提供)
- ② 構造化データ (.csv .pngなど)
- ③ メタデータ (.json)
- ④ 可視化データ (.jpeg, .png)

データカタログ
(書誌情報)

ARIMデータポータル (データマート)



https://nanonet.mext.go.jp/data_service/

データ共有
(ライセンス) 利用者

書誌情報を利用し、必要なデータを簡単に見つけることができます。
申し込みを受けた後、適切なデータセットがライセンス提供されます。

ARIMのデータカタログ

データセットの内容（一瞥）

DataSet データセット

クリア この条件で検索する

フリーワード検索

DOIあり/なしからさがす

- ☒ DOI指定なし
- ☐ DOIあり (0)
- ☐ DOIなし (42)

実施機関からさがす

- ☐ 物質・材料研究機構 (4)
- ☐ 東北大学 (3)
- ☐ 東京大学 (0)
- ☐ 名古屋大学 (1)
- ☐ 京都大学 (1)
- ☐ 九州大学 (2)
- ☐ 産業技術総合研究所 (3)
- ☐ 北海道大学 (0)
- ☐ 東京工業大学 (0)
- ☐ 量子科学技術研究開発機構 (2)
- ☐ 名古屋工業大学 (2)

新着順

表示件数 20件 表示順 登録日 降順

42件中 41~60件 < 1 2 3

Monochromated EELS of NiO

課題名: Monochromated EELS of NiO
課題番号: JPMXP1222NM1004
データ数: 4
実施機関: 物質・材料研究機構
登録日: 2023.10.27

EELS spectra of LiCoO2 and related materials for Li-ion battery

課題名: EELS spectra of LiCoO2 and related materials for Li-ion battery
課題番号: JPMXP1222NM1002
データ数: 10
実施機関: 物質・材料研究機構
登録日: 2023.10.26

42件中 41~60件 < 1 2 3



環境対応型超高分解能TEM

データカタログ

データセット名: Monochromated EELS of NiO

課題名: Monochromated EELS of NiO

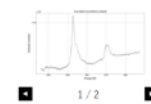
データセット登録者 (所属機関): MATSUNAMI, Shigeyuki (NIMS)
KIMOTO, Koji (NIMS)

課題番号: JPMXP1222NM1004
実施機関: 物質・材料研究機構

ダイレクトIN
カートIN

要約

EELS装置の校正に用いられるNiOのスペクトル。モノクロメーターを使用。
Standard EEL spectra of an NiO thin film, which is one of standard specimens for calibrating energy.



本報告書は「物質・材料研究のための先進電子顕微鏡」講義社 (2020). ISBN 978-4-065203866

キーワード・タグ

重要技術領域 (主): 量子・電子制御により革新的な機能を発現するマテリアル

マテリアルの高度化のための技術

重要技術領域 (副): 次世代ナノスケールマテリアル

横断技術領域: 計測・分析

マテリアルインデックス: 無機系材料・物質群 エレクトロセミコン

キーワードタグ: TEM EELS

データメトリックス

ページビュー: 210
ダウンロード数: 221

データインデックス

<https://doi.org/10.1038/nature06352>
登録日: 2023.10.27
エンバグ解除日: 2025.03.31
データセットID: 21836235-a5d5-46d0-b7ab-475dc3c4fd5a
データタイトル数: 4
ファイル数: 20
ファイルサイズ: 82.02MB

設置・プロセス

NIM-402: 原子力分析電子顕微鏡

成果発表・成果利用

論文・プロシーディング:
Koji Kimoto, Element-selective imaging of atomic columns in a crystal using STEM and EELS, Nature, 450, 702-704 (2007).
DOI: <https://doi.org/10.1038/nature06352>



大面積超高速電子ビーム描画装置

先端計測から最新微細加工のデータセットまで各種とりそろえています

データカタログのレイアウト

■ 世界的なデータカタログのデザインを参考に、データを探しやすく、より直感的に使える工夫へ。

サムネイルの複数表示

スライダー方式による複数枚の表示化

分類タグ・キーワードの設置

- ・重要技術領域(主・副)
- ・横断技術領域
- ・マテリアルインデックス
- ・キーワード

データセット名：Monochromated EELS of NiO

課題名：Monochromated EELS of NiO

データセット登録者(所属機関)：KIMOTO,Koji (NIMS)

課題番号：JPMXP1222NM1004

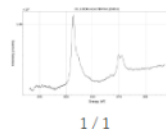
実施機関：物質・材料研究機構

ダイレクトIN

カートIN

要約

【【要約】】
EELS装置の校正に用いられるNiOのスペクトル。モノクロメーターを使用。
Standard EEL spectra of an NiO thin film, which is one of standard specimens for calibrating energy.
【試料および観察条件】
試料はカーボン膜上のNiO多結晶
Ni-L3ピークが852.75 eVとなるように装置(あるいはデータのエネルギー損失軸)を設定することが望ましい。
【参考資料】
木本浩司ほか「物質・材料研究のための透過電子顕微鏡」講談社(2020).ISBN 978-4065203866
Monochromator for TEM: Kimoto, Microscopy, 63 (2014) 337-344.
<http://dx.doi.org/10.1093/jmicro/dfu027>



キーワード・タグ

重要技術領域(主)： **マテリアルの高度評価のための技術**

重要技術領域(副)：

横断技術領域： **計測・分析**

マテリアルインデックス：

キーワードタグ： **TEM** **EELS**

データメトリックス

ページビュー： 522

ダウンロード数： 7

データインデックス



<https://doi.org/10.71947/ARIMJPMXP1222NM1004>

登録日： 2023.10.27

エンバーゴ解除日： 2023.10.31

データセットID：
21836235-a5d6-46d0-b7ab-475dc3c4fd5a

データファイル数： 4

ファイル数： 20

ファイルサイズ： 75.43MB

装置・プロセス

[NM-402：単原子分析電子顕微鏡](#)

成果発表・成果利用

論文・プロシーディング1：
Koji Kimoto, Practical aspects of monochromators developed for transmission electron microscopy, *Microscopy*, 63, 337-344(2014).
DOI:
<http://dx.doi.org/10.1093/jmicro/dfu027>

データセットDOIの新設

データセットDOIの記載項目を設置

機器IDによるクロスリンク

共用装置情報がクロスリンク

成果とのクロスリンク

成果発表・利用の記入欄設置。論文DOI入力による自動クロスリンク

機器利用者様（データ登録者）の成果普及（Visibility）の最大化を支援

お申込み方法



https://nanonet.mext.go.jp/data_service/page/registration.html

① データ利用約款への同意

「マテリアル先端リサーチインフラデータ利用約款」を確認する。

② ライセンス料金の支払い

令和6年度(2024年度)まで 試験的データセット利用は無償提供。
令和7年度(2025年度)よりライセンス料金の設定予定。

③ 会員登録の申込

事務局で内容を確認後に、アカウント(ID・初期パスワード)を発行。

④ データポータルサイトへログイン

https://nanonet.mext.go.jp/data_service/

⑤ 利用したいデータセットの検索

フリーワード検索の他、
機関、重要技術領域などのカテゴリごとに検索可能。

⑥ 利用したいデータセットのダウンロード申込

利用したいデータセットが見つかりましたら、
当該データセットの「カートIN」のボタンを押してカートに登録、
その後、下記の表示にある
「上記全てをダウンロード依頼する」ボタンを押して申し込み完了。

⑦ データの準備後に、ダウンロード

お申込みはデータポータルサイトから。まずは会員登録へ。
(来年度よりデータライセンスは有料予定)

来年度からはじまるARIMのデータ共用サービス



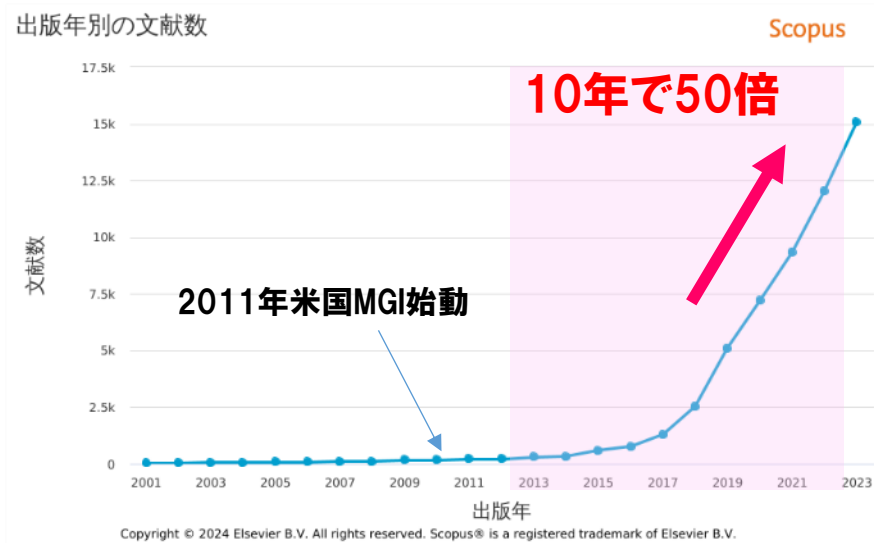
2025年度(令和7年度)から共用サービスを開始予定！

(2024年度は試行期間として限られたデータセットを無料でライセンスしています)

はじめに： このワークショップで伝えておきたいこと

世界を知る 材料科学・化学分野におけるインフォマティクス活用

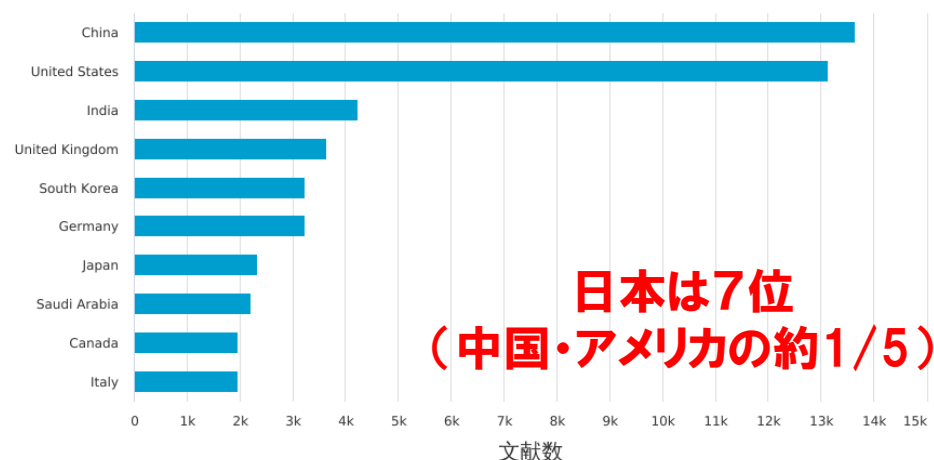
- 検索語: 「machine learning (機械学習)」を要約、タイトル、キーワードに含むもの
- 分野: 「材料科学 (material science)」および「化学 (chemistry)」
- 期間: 2001年～2023年



国/地域別の文献数

最大15か国/地域の文献数を比較する。

Scopus

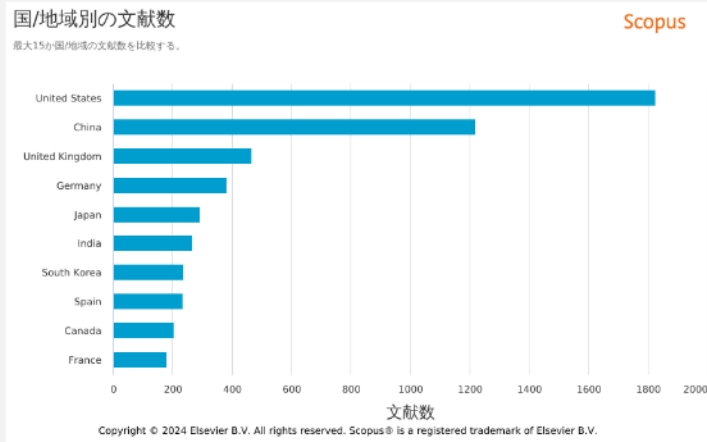
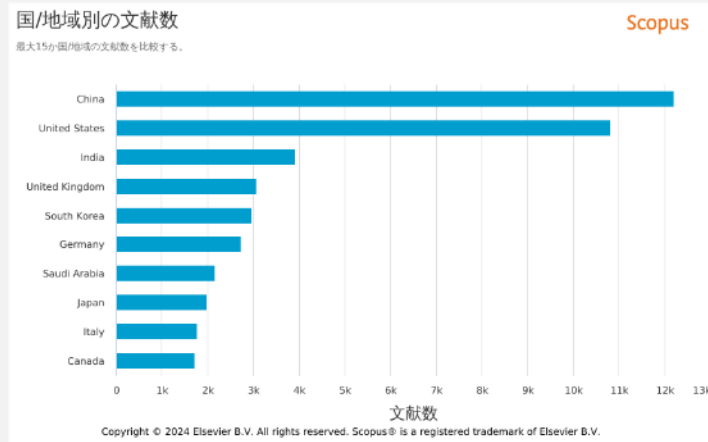


Copyright © 2024 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V.

- ・ 年間で15,000報の論文等が発表。
- ・ 材料科学分野・化学分野でデータ駆動型研究 (Informatics活用) は急増している

世界的にInformaticsの手法を研究開発に取り込むことは**当たり前**となっている

直近10年におけるプレーヤーの変化

	2014年～2018年	2019年～2023年
	<p>国/地域別の文献数 最大15か国/地域の文献数を比較する。</p>  <p>文献数</p> <p>Copyright © 2024 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V.</p>	<p>国/地域別の文献数 最大15か国/地域の文献数を比較する。</p>  <p>文献数</p> <p>Copyright © 2024 Elsevier B.V. All rights reserved. Scopus® is a registered trademark of Elsevier B.V.</p>
Top 3	アメリカ・中国・イギリス	中国・アメリカ・ <u>インド</u>
2nd 3	ドイツ、 <u>日本</u> 、インド	イギリス・ <u>韓国</u> ・ドイツ
Best 10	韓国・スペイン・カナダ・フランス	<u>サウジアラビア</u> ・ <u>日本</u> ・イタリア・カナダ

- アジアではインドと韓国の躍進
- フランス、スペインの陥落。代わってイタリアとサウジアラビアが台頭

【参考】直近10年におけるプレーヤーの年代別変化

	rank	2014		2015		2016		2017		2018		2019		2020		2021		2022		2023	
Class A	1	United States	139	United States	198	United States	273	United States	454	China	810	China	1851	China	2047	United States	2293	China	3341	China	4772
	2	China	102	China	191	China	227	China	436	United States	809	United States	1348	United States	1909	China	2111	United States	2644	United States	2956
	3	United Kingdom	35	United Kingdom	57	United Kingdom	64	United Kingdom	117	United Kingdom	196	United Kingdom	382	South Korea	545	India	733	India	1157	India	1553
Class B	4	Germany	18	Germany	36	Germany	55	Germany	84	Germany	185	South Korea	296	India	509	South Korea	720	South Korea	807	United Kingdom	904
	5	France	17	India	35	India	52	Japan	82	India	148	Germany	292	United Kingdom	493	United Kingdom	651	United Kingdom	749	Saudi Arabia	889
	6	Spain	15	Switzerland	34	Japan	43	India	70	South Korea	148	India	243	Germany	412	Germany	583	Saudi Arabia	746	South Korea	866
Class C	7	Switzerland	14	France	30	Australia	36	Spain	64	Japan	141	Japan	234	Japan	323	Japan	415	Germany	700	Germany	840
	8	India	13	Spain	30	South Korea	35	South Korea	53	Canada	106	Canada	216	Spain	290	Saudi Arabia	415	Japan	489	Japan	593
	9	Japan	13	South Korea	26	Spain	34	Canada	41	Italy	97	Spain	183	Canada	278	Italy	386	Italy	472	Italy	580
	10	Canada	12	Japan	23	Canada	33	France	37	Spain	97	Italy	167	Australia	241	Spain	368	Pakistan	439	Canada	531
Class D	11	Iran	12	Australia	21	France	23	Australia	34	France	82	Australia	164	Italy	228	Canada	340	Canada	436	Pakistan	462
	12	Russian Federation	7	Canada	19	Poland	23	Italy	33	Australia	75	France	133	Pakistan	203	Pakistan	309	Australia	361	Australia	424
	13	South Korea	7	Poland	17	Switzerland	17	Malaysia	30	Taiwan	64	Taiwan	111	Saudi Arabia	202	Australia	307	Spain	341	Spain	400
	14	Sweden	7	Italy	16	Iran	16	Switzerland	27	Russian Federation	59	Russian Federation	110	France	198	Malaysia	254	Russian Federation	310	France	376
	15	Australia	6	Taiwan	16	Italy	12	Russian Federation	26	Switzerland	58	Saudi Arabia	103	Russian Federation	198	Russian Federation	250	Egypt	308	Egypt	301
Class E	16	Austria	5	Iran	14	Netherlands	12	Taiwan	25	Netherlands	46	Malaysia	90	Taiwan	176	France	249	France	301	Russian Federation	298
	17	Italy	5	Brazil	13	Russian Federation	11	Iran	24	Poland	37	Pakistan	90	Malaysia	141	Taiwan	229	Malaysia	272	Turkey	296
	18	Malaysia	5	Russian Federation	12	Taiwan	11	Singapore	24	Malaysia	33	Switzerland	87	Iran	133	Brazil	190	Taiwan	250	Malaysia	291
	19	Poland	5	Sweden	12	Malaysia	10	Sweden	20	Sweden	32	Singapore	69	Switzerland	119	Switzerland	184	Turkey	246	Poland	249
	20	Turkey	5	Portugal	10	Belgium	9	Brazil	19	Pakistan	31	Brazil	67	Viet Nam	117	Poland	173	Iran	206	Taiwan	238

- 御三家： 米中英 → 中米印
- 続御三家： 独仏西（スペイン） → 英沙（サウジ）韓
- 2014年のトップ10入りしていたスイスとフランスが陥落。代わってイタリアがランクアップ
- サウジアラビアが2019年以降に大躍進のほか、ダークホースにパキスタン。

いろいろな要因

政 策

教 育

研究資金

産業構造
(経済対策)

人口動態

設備投資
(固定資産)

Software : python (オープンソース) の普及。最新のライブラリも入手可能。
Data : オープンデータの普及とメソッド開発の進展
Computer : 通常のCPU-PCでも十分に処理可。(いずれGPU-PCも低価格化へ)

(実験系研究のような) 設備投資のコストをかけずに資金の乏しい後進国でも最先端のインフォマティックス応用にキャッチアップできる。

これからの時代に生きるために

(大隅昇, 統計数理研究所・名誉教授)

探索的データ科学のススメ

「目的にあったデータの取得方法」が必要。そのためのデータ主導型の解析過程が必要

考え方:

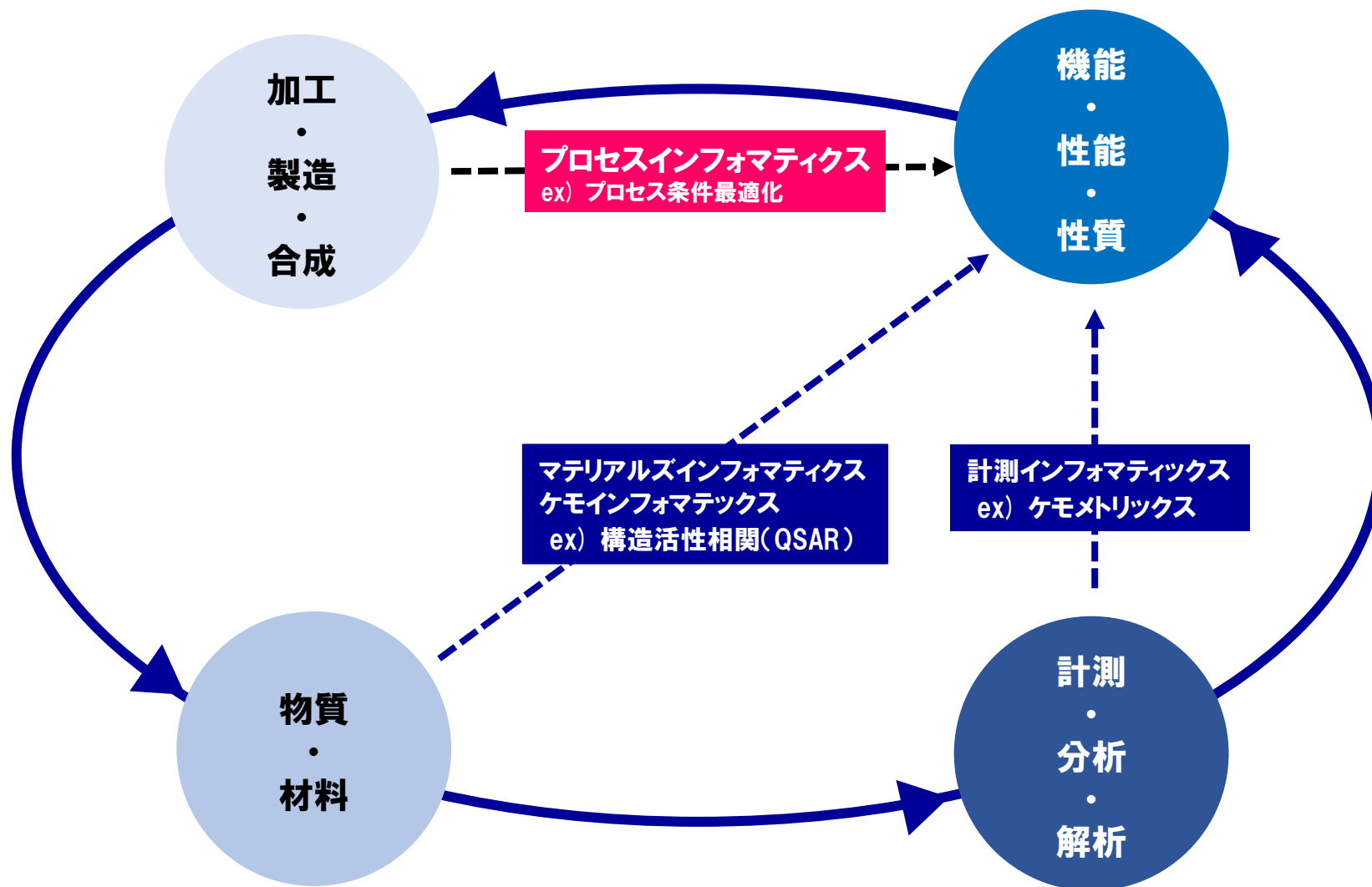
現象解析の本質は「データ」にある。データによる現象理解を前提として統計学、分類操作などを背景として統合的に現象解析をすすめる。

方法論:

- ① Experimental Design: データをどう計画的に取得するか
- ② Data Collection Mode: データを具体的にどう集めるか
- ③ Analyzing: 問題とする現象解析に適した解析法はどうあるべきか

マテリアル開発サイクルとデータ活用

- マテリアルの開発・評価ステージによって、データ構造化の設計は変化する。Informaticsを適用する技術分野ごとに要件を十分に吟味する。



【事例】茶のミネラル抽出条件の最適化

① 書誌情報（データカタログ情報）

論文名 (データセット名)	Mathematical optimization of multilinear and artificial neural network regressions for mineral composition of different tea types infusions.
著者名 (データセット作者名)	Durmus, Y., Atasoy, A.D. & Atasoy, A.F.
雑誌名 (課題番号、課題名)	Sci Rep 14, 18285 (2024). https://doi.org/10.1038/s41598-024-69149-1
発行日	2024年8月7日

② マテリアル情報

マテリアル名	茶葉
用途（産業利用）	ミネラル成分の分析（生産管理）

③ 装置利用情報

測定機器	誘導結合プラズマ発光分光分析法（ICP-OES）
メーカー（モデル）	Perkin Elmer（Optima 5300 DV）
試料名	4種類の茶葉：ブラックセイロン（BC）、ブラクトルコ（BT）、グリーンセイロン（GC）、グリーントルコ（GT）
測定条件	プラズマアルゴンガス流量（15 L/min）、補助アルゴンガス流量（1.5 L/min）、ネブライザーガス流量（0.75 L/min）、溶液吸収速度（1.5 mL/min）、および滞留時間（100 分）。すべてのテストは 2 回実施。

④ データセット基本情報

データセット要約	ブラックセイロン(BC)、ブラクトルコ(BT)、グリーンセイロン(GC)、グリーントルコ(GT)の4種類の茶葉について、3つの濃度（1%、2%、3%）で抽出したICP-OESの元素分析値。
ファイル拡張子	.CSV
データ構造	説明変数：Al、Ca、Cd、Cr、Cu、Hg、Fe、K、Mg、Mn、Na、Pb、Zn 目的変数：試料名（ラベル）
ファイル数	1

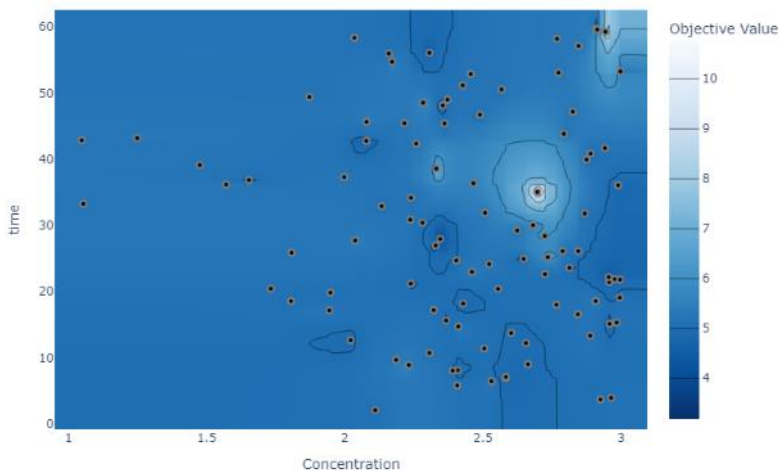
束縛条件：

- ・ Fe の抽出量はできるだけ多く、
- ・ Al の抽出量はできるだけ少なく、
- ・ Mn は特定の値で一定とせよ

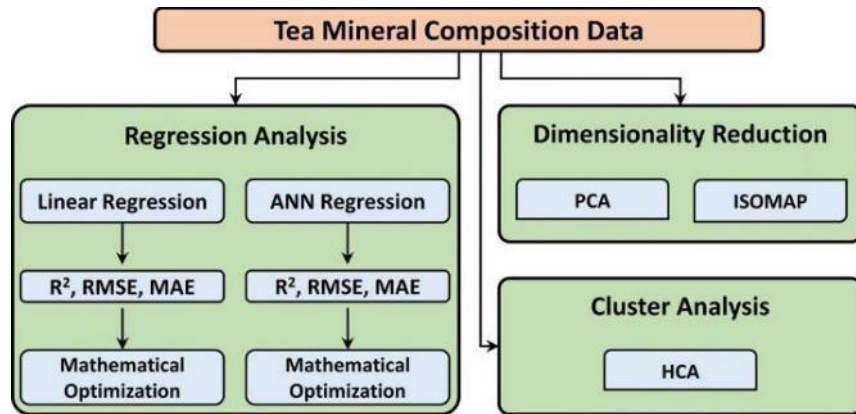
指示：

- ・ お茶の仕込み濃度と抽出時間を最適化せよ

Contour Plot



お茶のミネラル抽出のプロセスデータの最適化の方法



ミネラルの抽出量
の予測モデル
(scikit-learn)

束縛条件のもと
のベイズ最適化
(Optuna)

⑤ データ利活用

解析手法

茶のサンプル内のミネラル含有量を推定するための
① MLR と ANNによる回帰分析
② 主成分分析 (PCA) と等尺性マッピング (ISOMAP)、階層的クラスター分析 (HCA)による元素分析値のクラスタリング
③ Parzen Estimatorアルゴリズム (tree-structured Parzen estimator (TPE)) を用いた茶の抽出濃度・時間の最適化

ソフトウェア等

論文ではMinitab → Python (scikit-learn, keras, Optuna)

サンプルコード

- ① MLR, ANNによる回帰分析
https://colab.research.google.com/github/ARIM-Usecase/Example_2/blob/main/1_ML_Code-1.ipynb
- ② PCA, ISOMAP, HCAによるクラスタリング
https://colab.research.google.com/github/ARIM-Usecase/Example_2/blob/main/2_DR_Code-2.ipynb
- ③ TPEによる抽出条件の最適化
https://colab.research.google.com/github/ARIM-Usecase/Example_2/blob/main/3_TPE_Code-3.ipynb

ワークショップを終えたあと、是非、試してみてください

ARIMのデータ人材育成カリキュラム

①初級者向け

データ活用講座
データ構造化オンライン学習

3回のオンライン方式

開講：初夏
定員：100名程度
対象：**一般向け**
(ARIM機器利用あり)

②中級者向け

ARIMアカデミー
データ構造化ワークショップ

3日間のオンサイト方式

開講：秋～冬
定員：20～30名
対象：**一般向け**

③上級者向け

データ活用講座
インフォマティクス学習

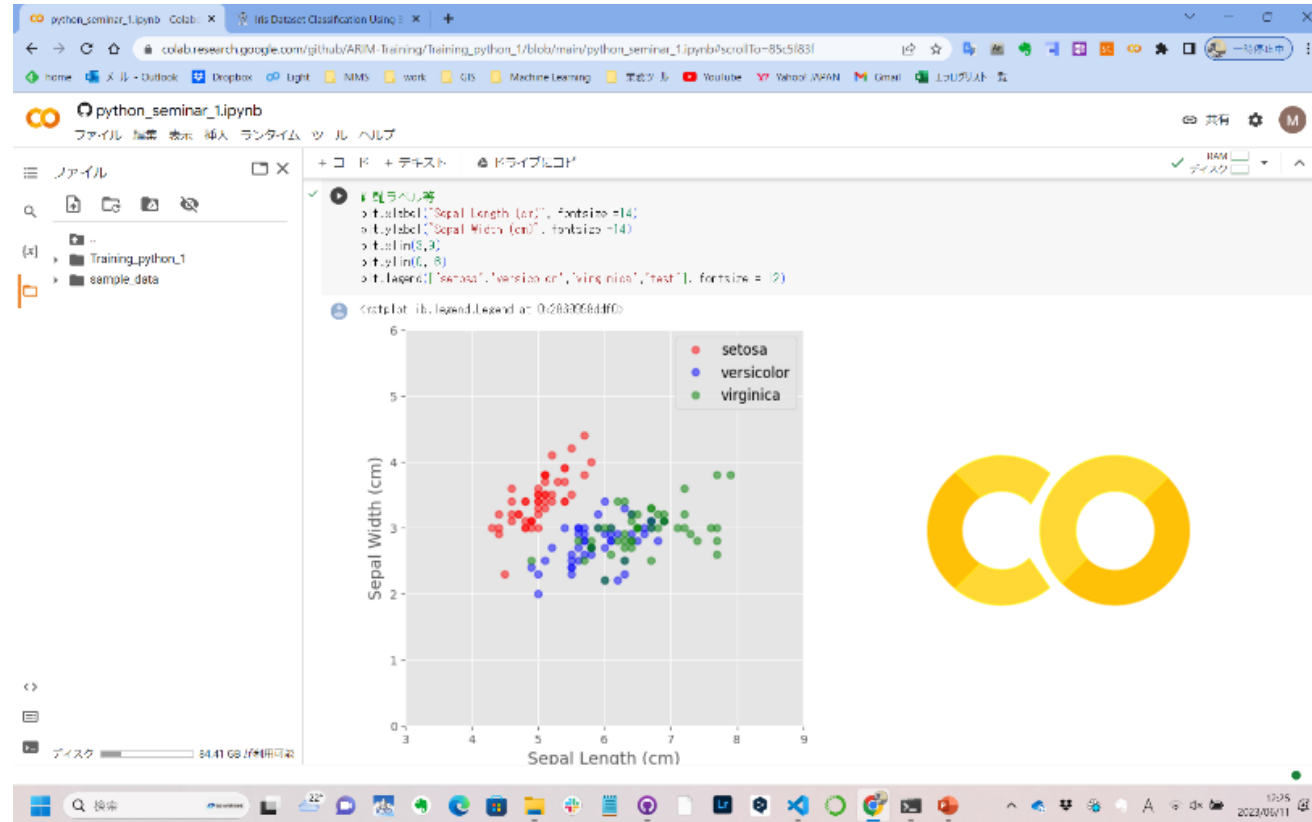
1日オンサイト方式

開講：不定
定員：不定
対象：**事業内外の希望者**

ARIMアカデミー データ構造化ワークショップ

	セッション名	学習内容
Day1 10:00～	10:00 - 10:30 1日目 イン트로ダクション	—
	10:30 - 11:00 グループ自己紹介	研究・業務・学業等の紹介
	11:00 - 12:00 機械学習 入門編	機械学習の概要
	12:00 - 13:00 お昼休憩	—
	13:00 - 14:00 機械学習 実践編	scikit-learnによる機械学習の実践
	14:00 - 15:40 機械学習 グループワーク	—
	15:40 - 16:40 機械学習 グループワーク発表会	—
Day2 9:30～	09:30 - 09:40 2日目 イン트로ダクション	—
	09:40 - 10:30 ハイパーパラメータ 入門編	ハイパーパラメータの基礎知識
	10:40 - 12:30 ハイパーパラメータ 実践編	Optunaによるチューニングの実践
	12:30 - 13:30 お昼休憩	—
	13:30 - 16:40 <特別講演>	【ご講演内容】
	奈良先端科学技術大学院大学 データ駆動型サイエンス創造センター	Part I データ駆動化学の発展の歴史と展望
	船津公人センター長	データ駆動化学はどのような歴史を辿りいまに至っているのか。 黎明期から現在までの取組みのマイルストーンを見つつ今後の動きを展望する。
Day3 9:30～	09:30 - 09:40 3日目 イン트로ダクション	Part II データ駆動化学が導く研究・開発・生産のパラダイム変革～リサーチトランスフォーメーション (RX) サイクルの実装～
	09:40 - 11:40 ハイパーパラメータ グループワーク	データ駆動化学における主要なカテゴリーでの取組みを紹介し、それぞれの取組みにおけるデータ、情報の有機的連携の姿としてNAIST/DSCで進められているRXサイクルを実装するRXプラットフォーム構築へと話を進めたい。
	11:40 - 12:10 ハイパーパラメータ グループワーク発表会	—
	12:10 - 12:30 中間クロージング	—
	12:30 - 13:30 お昼休憩	—
	13:30 - 15:00 Pythonプログラミングの勘所 [任意参加]	関数、クラスの上手な使い方など
	15:00 - 15:30 アンケート・クロージング [任意参加]	—

Colabによるオンライン演習



Googleアカウント(個人用でも何でも)を取得しておいてください