マテリアル計測・プロセスのためのデータ設計およびデータ構造化

2021年6月16日　JOSS2021　D1 研究データインフラ技術者＆研究者座談会

国立研究開発法人物質・材料研究機構

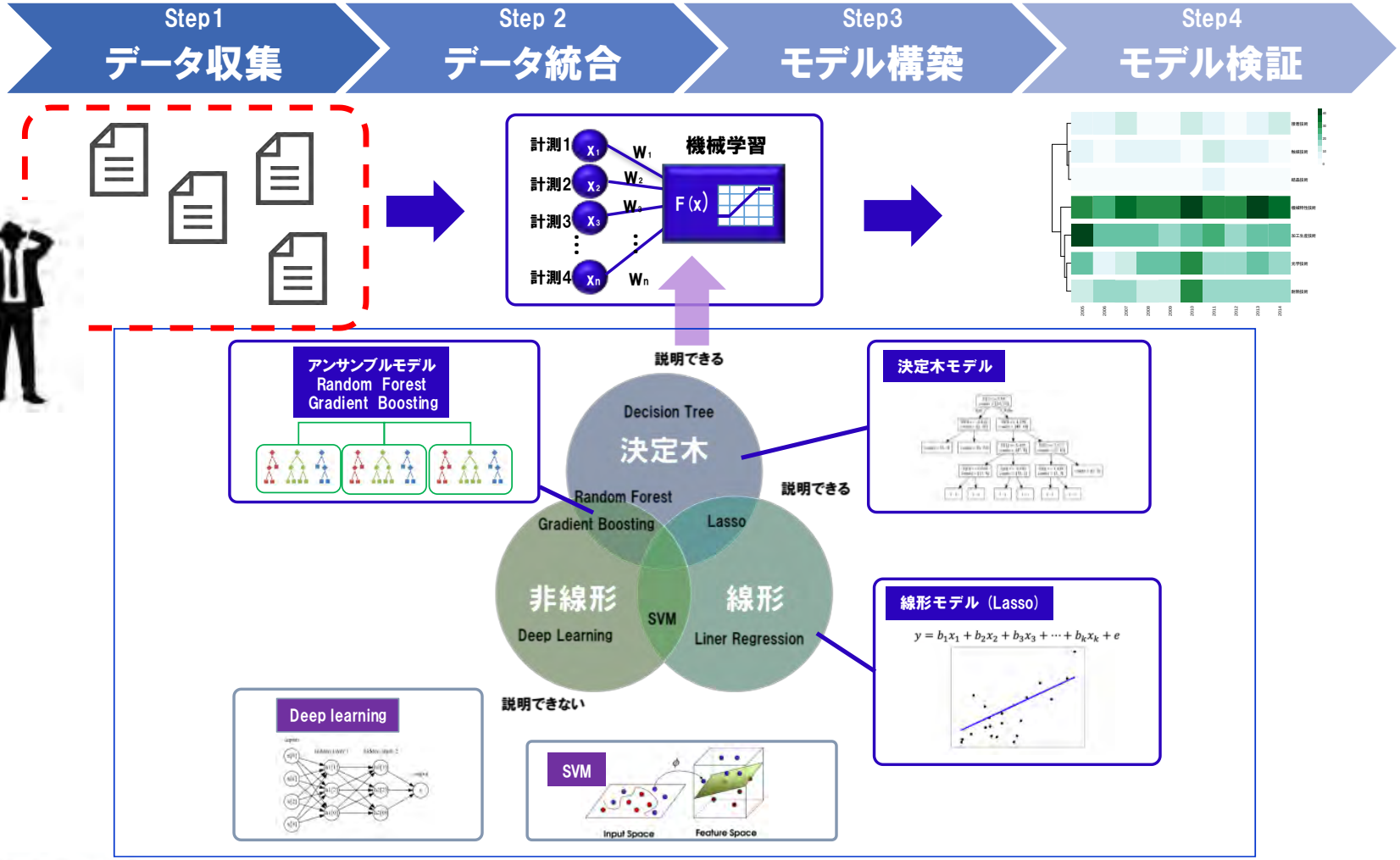
統合型材料開発・情報基盤部門

松波　成行

1. データ収集にかかる背景と課題

データ駆動型研究の進展と課題

- データ駆動型研究のトータルの作業工程の中で**川上側のデータ収集が最も律速**となっている
- データ収集を効率的に行えるようにするためには，・・・．



課題分析： データ収集ができない主な理由

■ 現状は依然としてデータの収集段階が最も障壁となっており，その理由の多くは，以下の事由の組み合わせによっている．

1. データそのものが測定装置本体や個人のローカルPCに保管されており，データの所在が分散している．そのため，データの所在を人に聞いて回らないと集められないこと．

2. 測定装置の多くはウイルスによる誤動作やOSのアップデートによる測定の中断などのトラブルを回避するため，ネットワーク環境から切り離したスタンドアロンで運用されている．そのため，データの取り出しや受け渡しはUSBなどの外部記録媒体を人手が介在していること．

3. 測定データの素性が不明であること．例えば，新規材料の開発においては試料情報が略語表記やコードネーム化されている．また，計測や分析における前提条件に関する付帯情報がないために，測定当事者以外の第三者による再利用・再解析が難しいこと．

4. 計測データが出力状態のまま（rawデータ）となっていること．特に出力がメーカー固有のバイナリー形式の場合には可読化のためのデータ変換に難航することのほか，記載内容もメーカー固有の計測パラメータの語彙の意味が解読できず，データ駆動にかけるためのデータ構造化が難しいこと．

取り組み： 測定データのデータセット化とハイスループット化

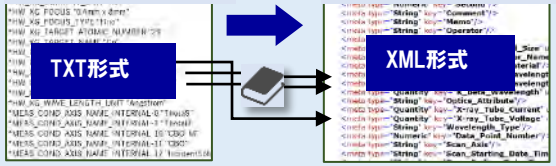
■ データ取得から前処理・解析を経て蓄積までを一気通貫 → 「ハイスループット化」 ＝省時間化
■ データ蓄積は利活用しやすいようにデータ構造化 → 「データセット化」＝知財化

材料計測のハイスループットな収集とインフォマティクス

データ蓄積

メーカー・検出器依存の装置出力

主要パラメータの抽へのメタ情報付与



構造化

XML or json
REST API

データセット化

収集

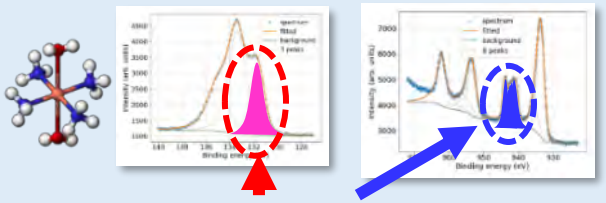
蓄積

処理・解析

IoTデータ転送

省力化

データ解析（例：スパースモデル解析）



自動でピーク分離される解析データ

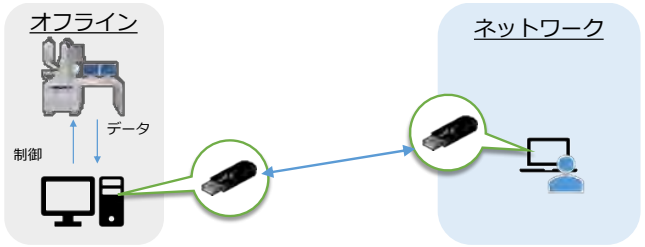
省時間化

2. IoTによるデータ収集システムの概要

IoTシステムの特徴

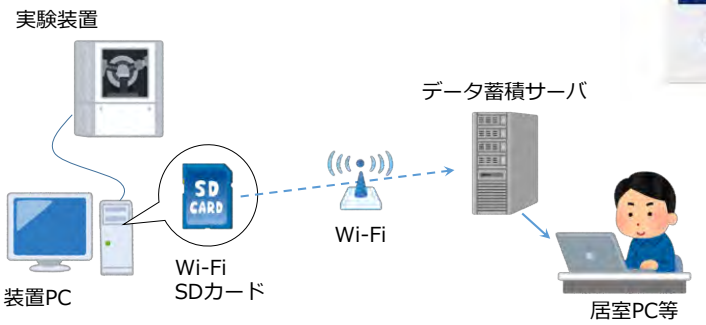
■ 対象機器： スタンドアロンで運用される計測・分析装置やプロセス装置の制御パソコン向け
■ 方　　法： 通信セキュリティを施したWiFi付きSDカードを活用

これまでの方法



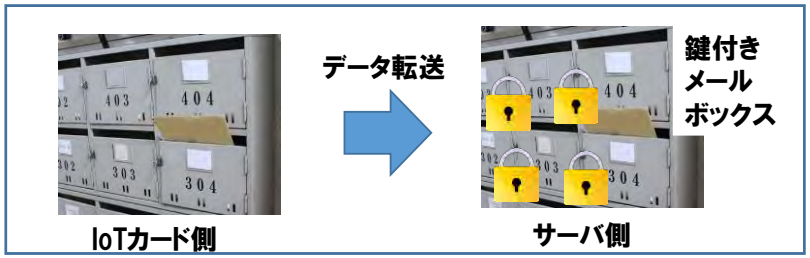
- OSの自動アップデート等で長時間測定が中断される！
- OSが新しくなると、装置の制御プログラムが更新できない！

本システム



特　　徴：

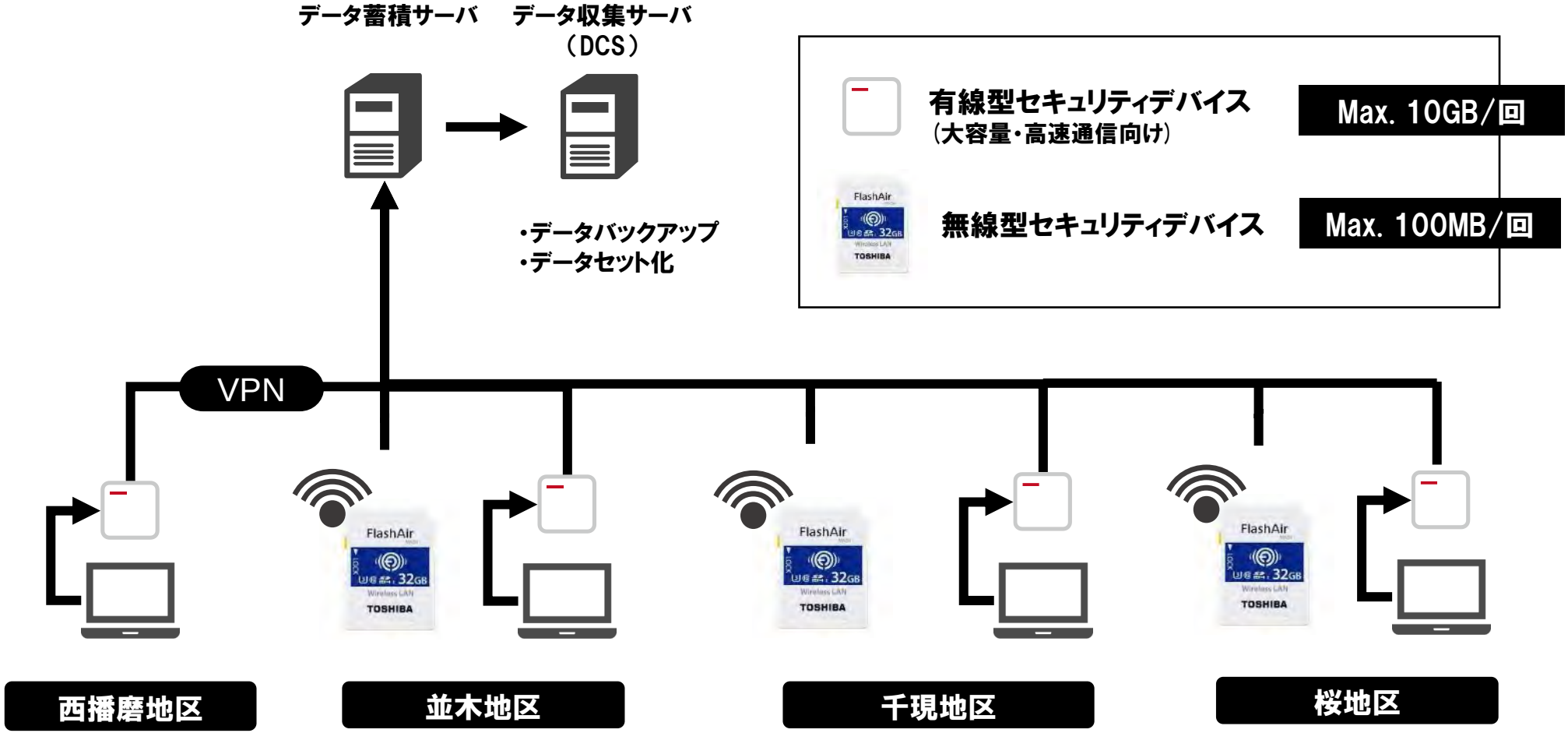
① 情報セキュリティを高めるアップロードのみの通信制御によるデータ転送
② データの秘匿性を確保する私書箱方式のデータの振り分け機能
③ 小型なので大きな工事が不要。誰でもが簡単に増設や設置が可能



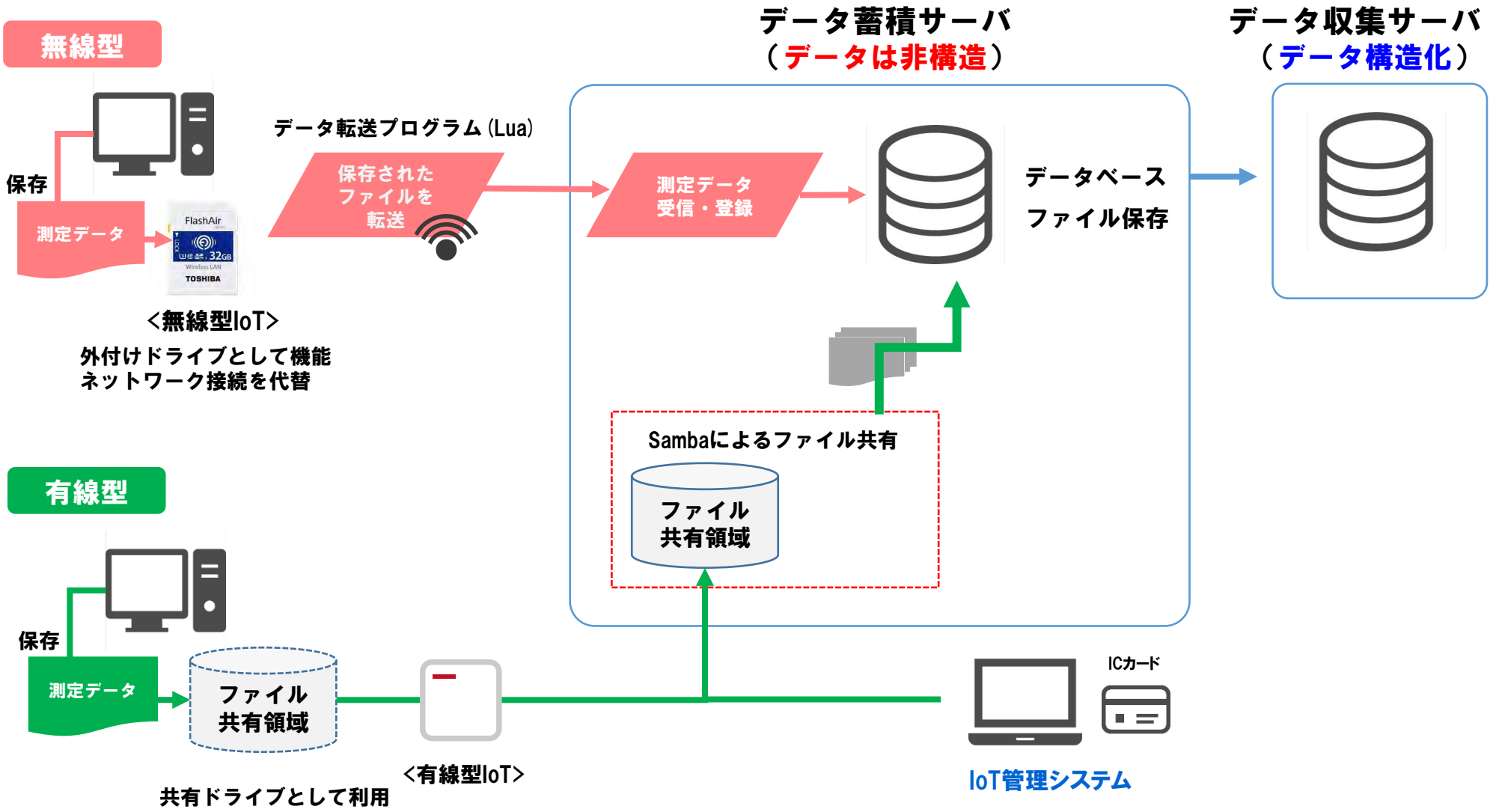
一般の産業用IoTサービスにはない、NIMSならではのシステム設計

IoTシステムの構成概略図

■ セキュリティを確保した**有線型と無線型のIoTセキュリティデバイス**による統合的なデータ収集システムを開発済



データ転送における構成図



3. データ設計の考え方

機器利用とデータ利用のワークフローの高度化

■ 作業効率化： 測定rawデータを即時にワンストップでデータ処理＆解析をフィードバック．
■ データセット利用： データセットを使って最新研究に．オリジナルの機械学習モデルも実装可能．

研究者の機器

データ転送

IoTセキュリティデバイスを装着が前提

ドラッグ＆ドロップのみ
お手持ちのPCからDropboxの感覚でドロップ！　非ネットワークもOK（IoT設置が条件です）

NIMS/DPFC

Python

前処理 or データ解析

ユーザーの自作したAIモデルも実装できるよ♪

定型のデータ構造化(前処理)や自動解析のコード実装
Pythonのコードを実装し，転送データをすばやく前処理，もしくはデータ解析を実行

機器利用者向け

速達

データ処理・解析速報

ダウンロードOK
知りたい結果を即時にお届け

データ利用者向け

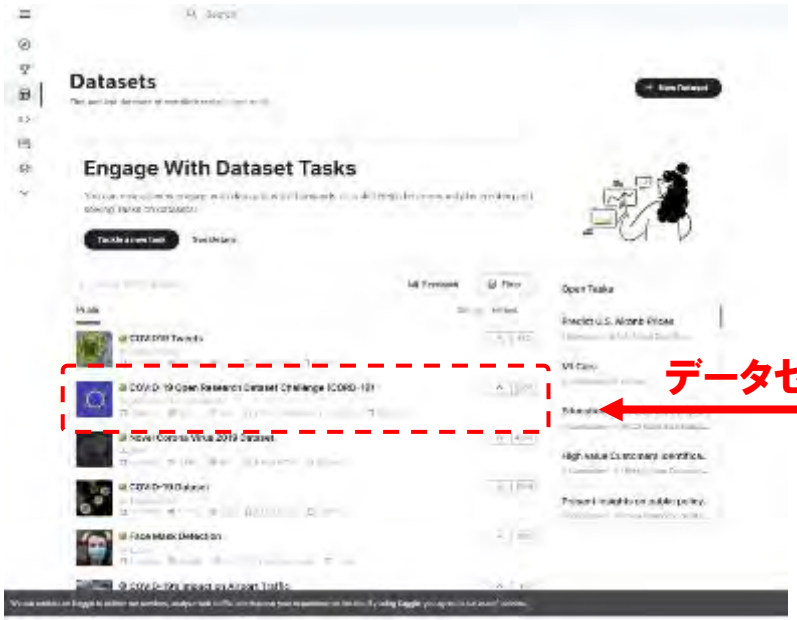
データセット

データ構造化
データセット化し教師学習などのデータに！

データセットとは

■ 特定の目的やテーマに沿って収集されたデータ群.
1） 統計処理や機械学習等にも利用しやすいように**データが構造化**された状態となっている.
2） データの属性や意味がわかる**データカタログ**を備える.

kaggle （有名なデータセットの例）



データセット単位

NIMS (Research Data Express)

目的とするデータ群（データセット）がすぐに閲覧/ダウンロードが可能です

データ構造化の意味

1. 統計学的・データ科学的な見地

- ✓ 説明変数と目的変数を明示的に定めること
- ✓ 説明変数(探索空間)は実験のログ，グラフ，もしくはスペクトルから特徴量を数値化すること

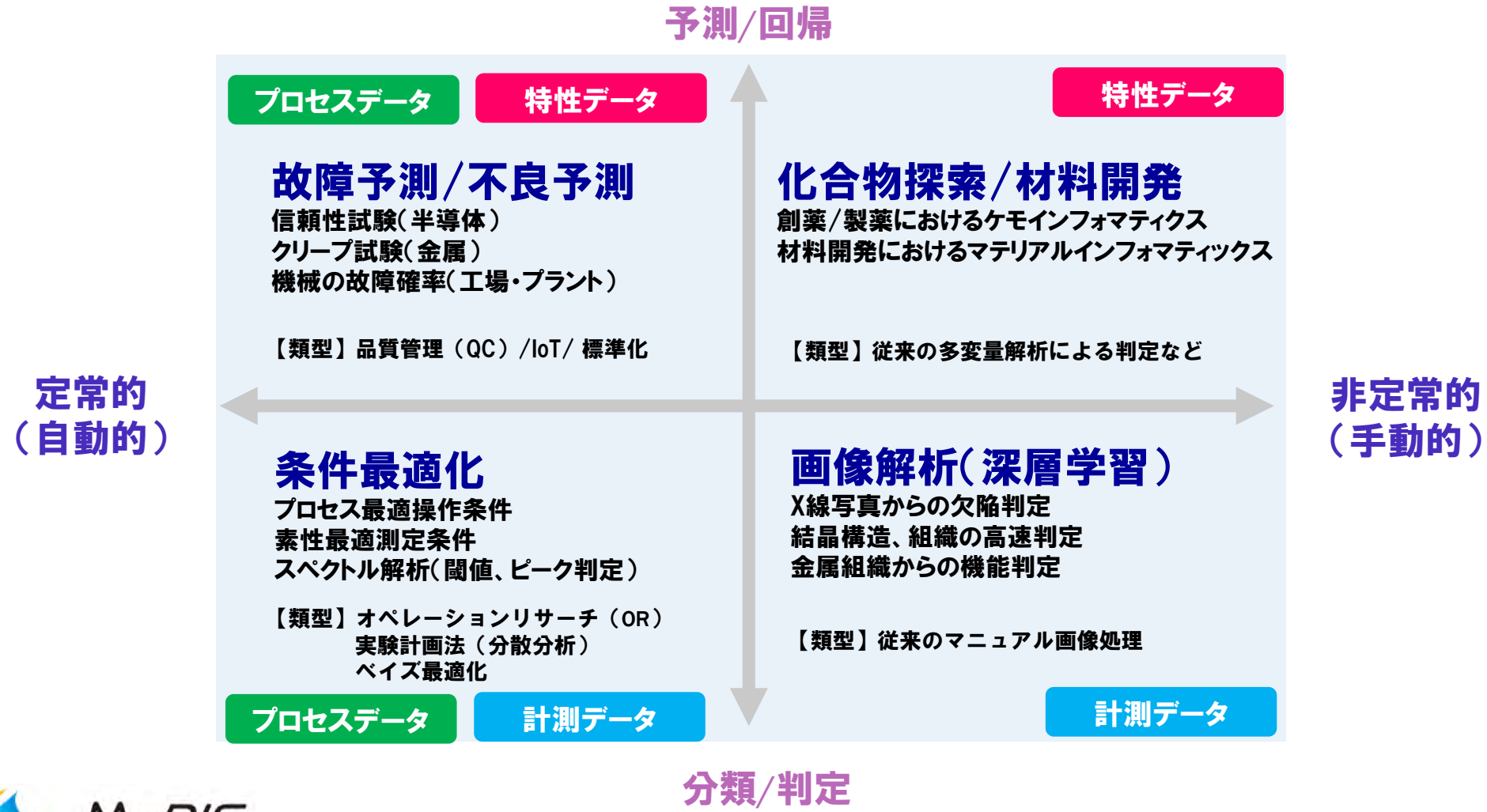
2. データベース工学的な見地

- ✓ 数値データ部はソフトウェア互換的の高いcsv形式に変換をかけること
  (多次元の場合には，hdf5形式，Net-cdf形式などの可読性のあるバイナリー化にすること)
- ✓ 実験(プロセス・計測)パラメータをXMLスキーマを定め，DBへ登録できること
- ✓ データの意味が直感的にわかる参照図(サムネイル)を付すこと

これらをできるだけ人の手を介さずに自動的にデータセット化すること

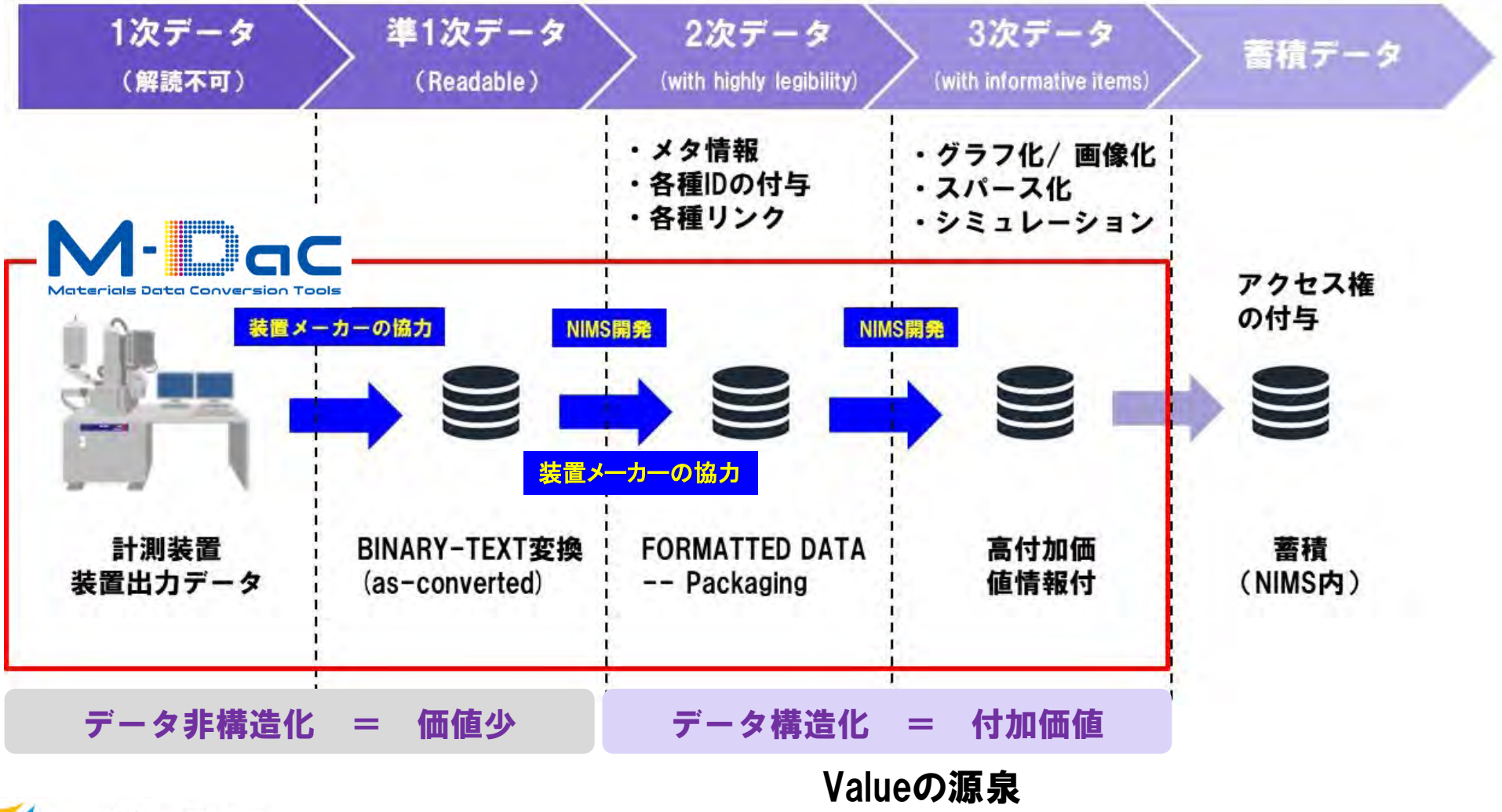
データセットの活用形態

■ データ構造化された「データセット」は，データ駆動型研究へもたされ，新たな解析手法の礎となる．
→ 開発された手法が現場の判断を高度にアシスト．省時間化で精度の高い結果をFB．



データ蓄積： 計測データの可読化と高付加価値化のフロー

■ 非構造化データから構造化データへ変換することで付加価値をつける



データ設計： インフォマティクスのためのデータ構造化

悩み： 本丸のデータ処理までの前処理/定形処理に膨大な作業工数がかかっている
→ 一般に機械学習(AI)にかかる検討の8割は前処理に割かれている

前処理（クレンジング）

- データ変換 （バイナリー変換、次元変換）
- データ成形 （並び替え、統合、抽出、分離）
- データ変形 （単位変換、規格化、標準化）
- 欠損値・外れ値・重複データ処理

最適化

- Fitting　Fittingパラメータの最適化
- 信号処理　ピーク分離、ピーク位置、ピーク面積
- 画像処理　二値化、粒界検出、物体検出
- 補間　線形補間

分類判定・集計

- 欠陥判定、エラー判定
- 類似度判定、パターン分類、クラスタリング
- ラベル化 （正例化、負例化）
- 異物カウント、分級化集計

材料の機械可読化

・材料(組成)の機械可読化

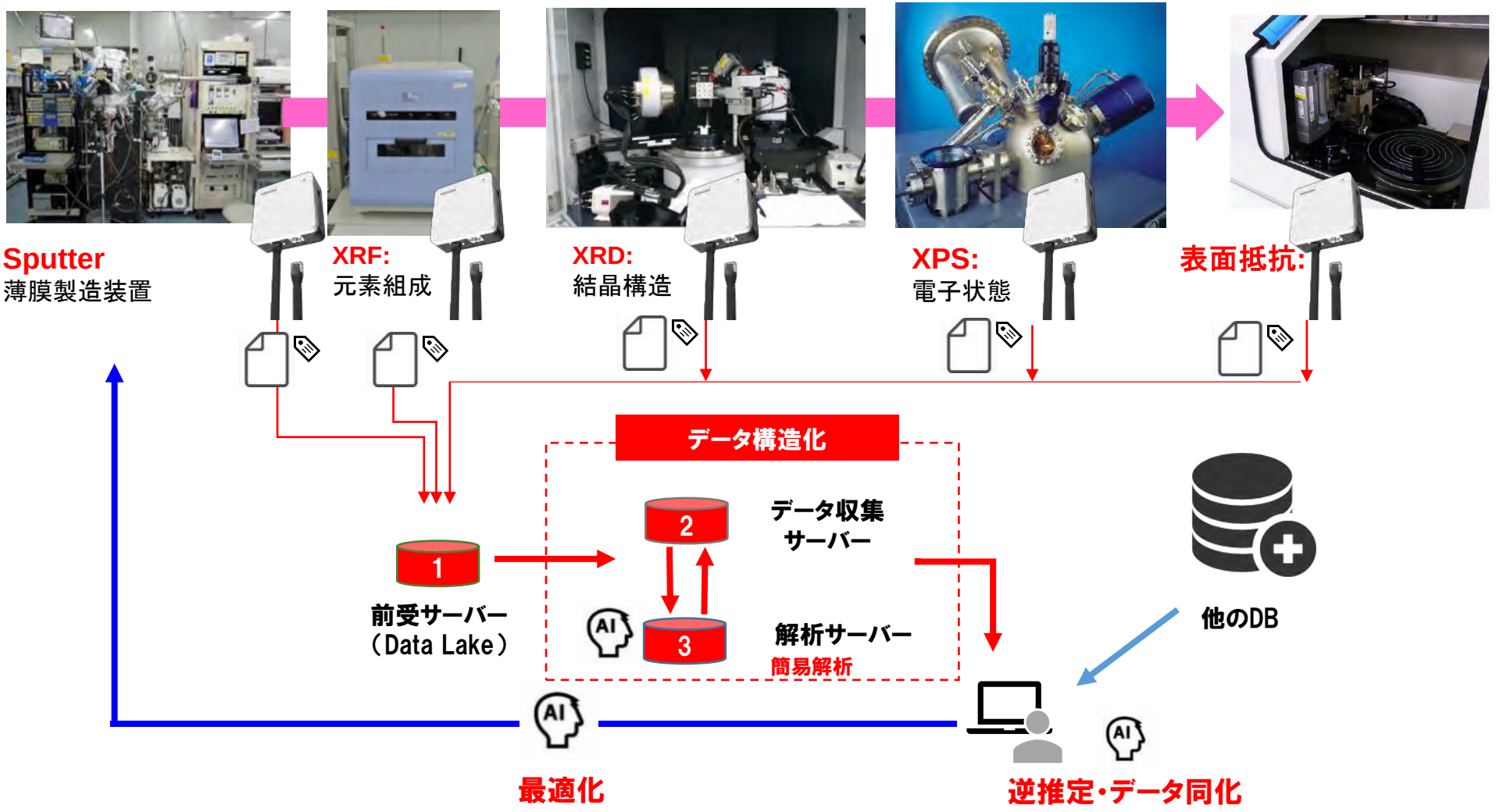
ex) 有機材料ではSMILES化

・組成からのFingerPrint化

データ収集サーバ(DCS)

4. 半導体薄膜評価でのデータ設計事例

IoTシステムによるラボ装置のデータ構造化

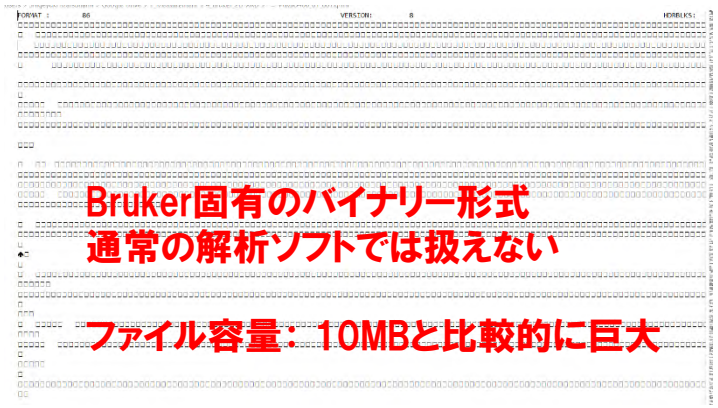


IoTセキュリティデバイスを装着することで，各種プロセスや計測機器のデータをセキュアに，そして効率的にデータセット化して蓄積することができます．

【事例1】 二次元XRDのシーケンスファイルの一次元自動積分

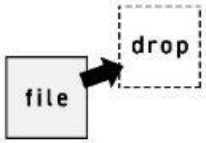
要　望: 二次元XRDデータから一次元積分を自動化したいが，メーカーのソフトでは対応できない．
適　用: pythonのスクリプト作成によりIoTで投入後に自動化完了．複数のシーケンスデータの連続処理も可能．

rawデータの種類 （gfrmファイル）



Bruker固有のバイナリー形式
通常の解析ソフトでは扱えない

ファイル容量：10MBと比較的に巨大

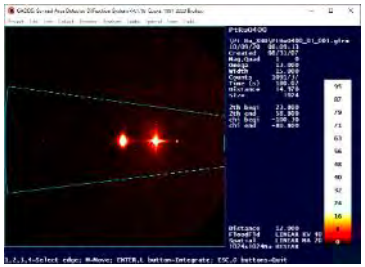


- binary⇒ txt化
- 一次元積分
- シーケンスファイルの連続処理

[redacted]速達[redacted]

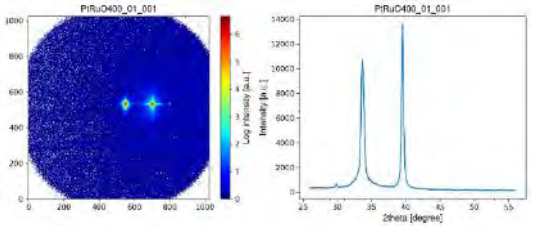
これまで

メーカのソフトで処理．自動化処理に難
複数処理にも対応が不可



適用後

IoT投入後10秒程度で処理完了．
複数処理も対応可



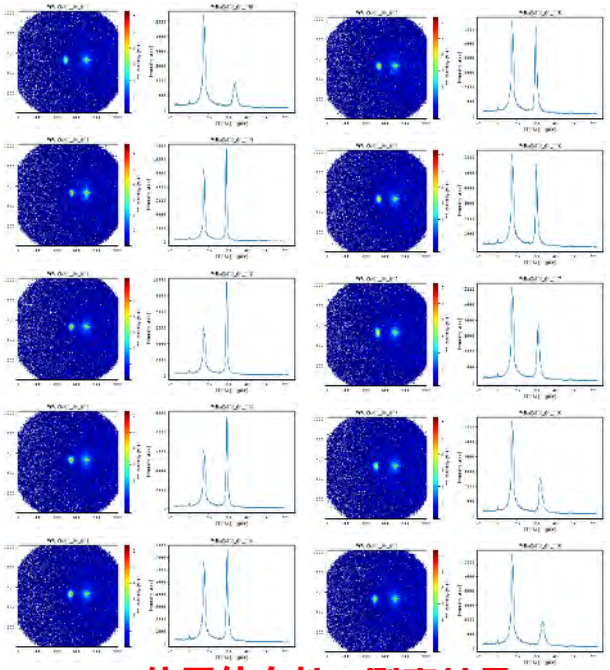
二次元 ⇒ 一次元積分を自動化

機器利用者においては

- 定型的な作業について大幅な作業時間の低減が見込めます
- 制御PCにバンドルされた解析ソフトによるオンサイトの作業から解放されます

1次元積分のマッピング（ライン測定のヒートマップ）

要　望：傾斜勾配のあるスパッタ膜の位置依存性を可視化．基板の組織状態の情報すぐに確認したい．
適　用：バイナリーデータから一気通貫化し，ヒートマップとして可視化．これまで1日がかりの作業を1分で速報．

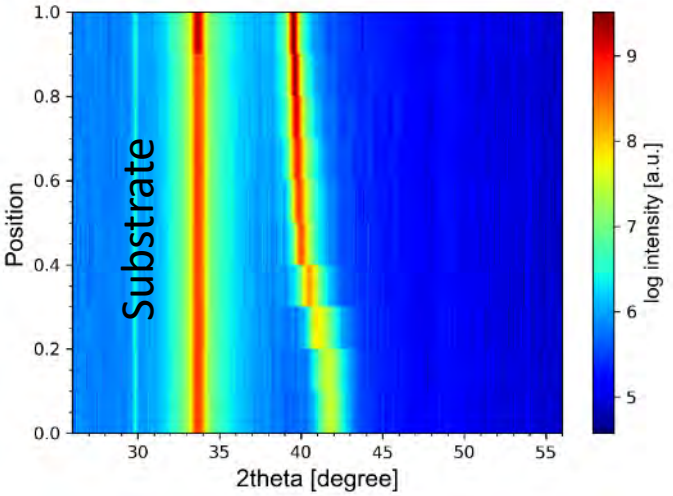


位置依存性の測定結果
（10ファイルで1分程度）

python

・1D積分値をマージ
・ヒートマップ化（補間図）

[redacted]速達[redacted]



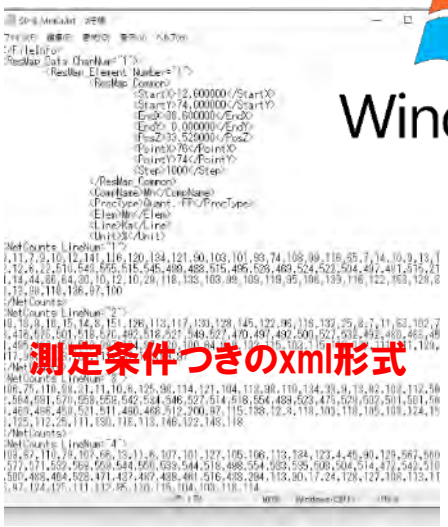
10点の計測からヒートマップ化
IoTのデータ投入から1分程度で描画完了

繰り返して行うようなデータ処理は，自動化で大幅な業務効率（時間短縮）が図れます．

【事例2】 XRFの組成二次元マッピング （20mm角基板編）

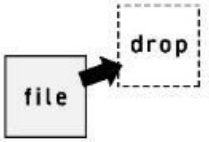
要　望: 機器が古く(Win 7), メーカ固有のソフトは非ネットワーク化の孤立PCの作業で自動化できない.
適　用: 瞬時に解析結果がフィードバックされ, 試料の出来を確認できる.

rawデータの種類 （xmlファイル）



測定条件つきのxml形式

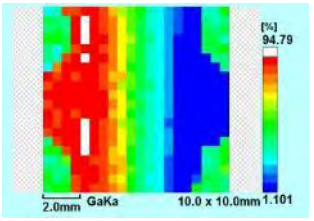
Windows 7

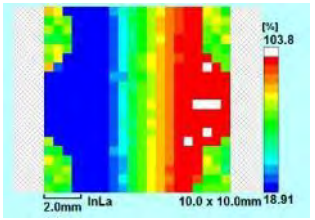


・xmlからの必要数値部の抽出
・ヒートマップ化(補間図)

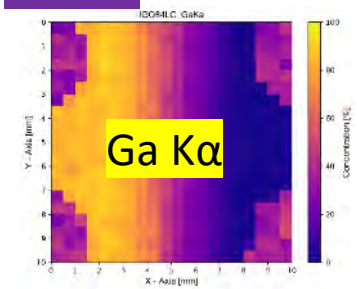
速達

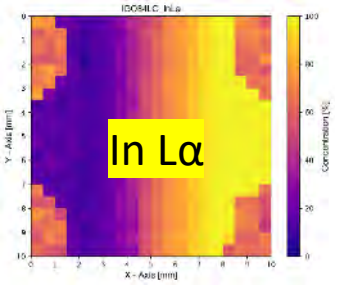
これまで　メーカの処理ソフトに依存
自動化ができていなかった





適用後　IoT投入後に10秒程度で処理完了





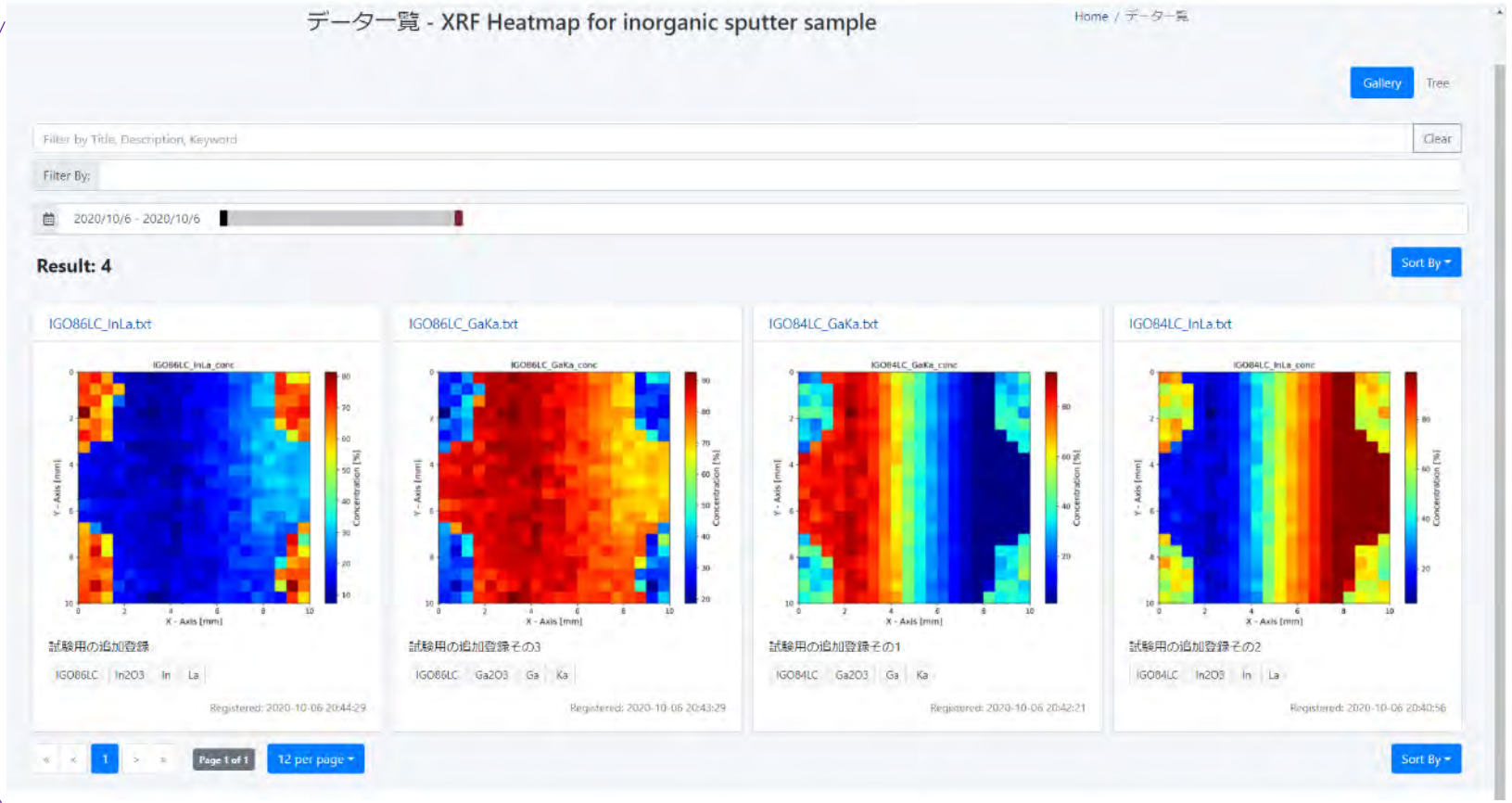
20mm角の組成傾斜スパッタ膜の特性X線

古いOSでもバージョンを問わず, データを活用・再利用ができるようになります.

MaDIS
NIMS MATERIALS DATA and INTEGRATED SYSTEM

データセットの登録状態　（Gallery形式）

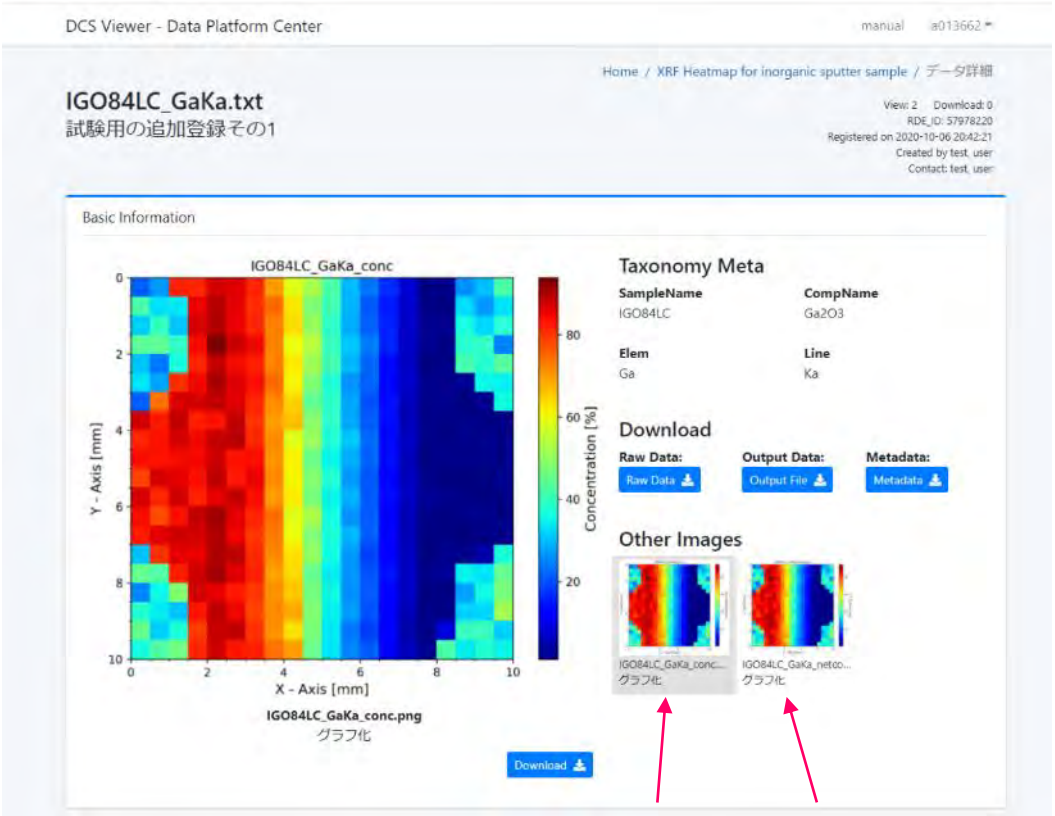
■ IoTで投入された単位ごとでファイル(データ)が構造化され，登録/表示される
■ 同一のフォーマットで系統的に蓄積されるため，長期的な傾向把握も可能となる



自動的に構造化されてデータベースに蓄積されてゆきます

データセットの登録状態（詳細画面）

■ メタデータ(メタ情報)の9割は計測ファイルから自動的に読み取り，取得される．
（読み取りのためのXMLスキーマは事前に準備し，そのスキーマに沿ってpythonで読み取る）

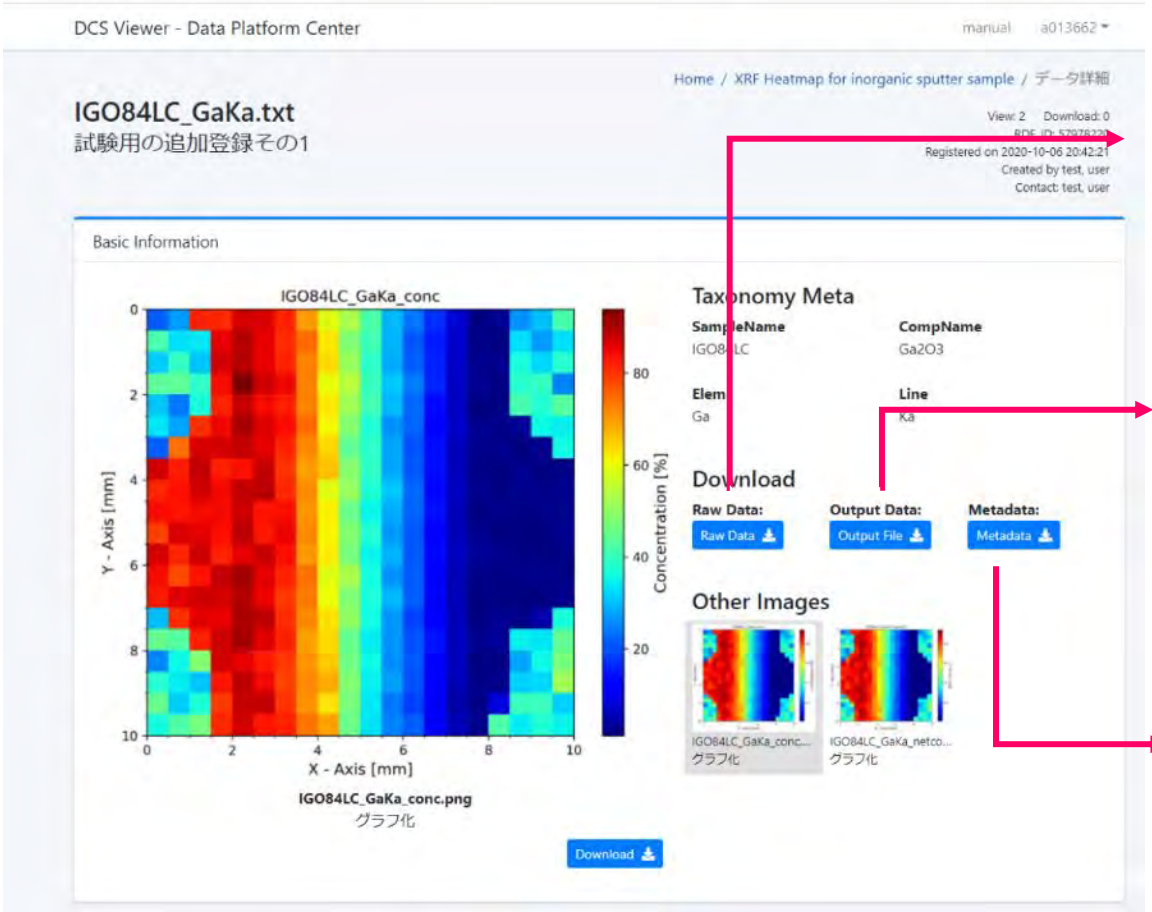


組成分率　強度

共通メタ：　ファイル名や装置名，資料IDなど．運用側での管理情報も含む
主要メタ：　主には計測ファイルに含まれる計測パラメータとデータ送付状の固有情報

データセットのダウンロード（詳細画面での場合）

- 入力データのほか，構造化された成形データや実験記録(メタデータ)をダウンロード可能
- ダウンロードできる対象は指定したグループ内/組織内で選択が可能



Raw Dataボタン

- IoTで投入されたrawデータ

OutPut Dataボタン

- 座標×濃度分率のcsv
- 座標×強度のcsv
- 可視化画像(png ファイル)

MetaDataボタン

- 共通メタ・主要メタのcsv
- (後述)データ送り状の記載項目付き

データ利用者には使いやすい形で構造化されたデータを提供することができます．

データカタログについて

- 記述項目は『**データセットの概要**』，『**データセットの諸元**』，『**計測メタデータリスト**』からなる
- どのような**公開方式**にしたいかなどのライセンスもここで決めておく．

公開方式

DPFC-DCS 公開データカタログ

（機能性材料研究拠点（RCFM）ナノ電子デバイス材料グループ内限定）

作成日：2020-09-14

■データセット概要

データセット名	XRF Heatmap for inorganic sputter sample
概　要	無機半導体の組成傾斜がある主に二元系スパッタ薄膜（X:基板 X 方向、Y:基板 Y 方向、Z：カウント数）の二次元ヒートマップ図。
作成者	長田貴弘
公開範囲	機能性材料研究拠点（RCFM）ナノ電子デバイス材料グループ内限定
ライセンス	論文前，公開前のデータにつき作成者の許可必要
言　語	日本語
使用装置	島津製作所　XRF　μEDX-1400

■データセット諸元

データ数	XRF（μEDX-1400）に設置の IoT より都度追加される
データの種類	XRF 二次元マッピング測定
raw データ属性	1. **拡張子**：　txt 2. **ファイルサイズ**：　10～20KB 3. **欠損値**：　なし（特に定められていない）
raw データのデータ構造	● ファイル内部は XML 構造表記がなされている。 ● ['FileInfo']，['ResMap_Common']，['ResMap_Element']の３つのブロックからなる。 ● ['FileInfo']は計測の書誌情報の項目，['ResMap_Common']はマッピングを行った基板の座標情報，['ResMap_Element']は蛍光 X 線の計測データが格納されている。 ● 「'ResMap_Element'」の配下には，「NetCounts」のサブブロックには正味の強度，「Conc」には濃度分率に変換された百分率がしめされている。
output ファイル属性	● raw データからは xrf_heatmap.py を通じて以下の６つのファイルが出力される．欠損値処理は特におこなっていない。 -- csv 形式 -- ①ファイル名_netcount_original.csv ファイルサイズ：　～10KB ②ファイル名_netcount_convert.csv ファイルサイズ：　～10KB ③ファイル名_conc_original.csv ファイルサイズ：　～10KB

v.2020912

■メタデータリスト

Viewer における英語語彙は出力ラベルメータの表記をそのまま転載している．日本語語彙は DPFC で任意に定めたものであり計測メーカーの了承の語彙表記ではないことに留意の上，計測パラメータの理解の参考とすること．

	出力計測パラメータ	日本語語彙	英語語彙	単位
1	['FileInfo']			
2	SampleName	サンプル名	SampleName	
3	AnalType	解析手法	AnalType	
4	Comment	コメント	Comment	
5	ParFLN	解析グループ	ParFLN	
6	MeasDateStr	測定日時	MeasDateStr	
7	Operator	オペレータ	Operator	
8				
9	['ResMap_Common']			
10	StartX	開始位置 X 座標	StartX	mm
11	StartY	開始位置 Y 座標	StartY	mm
12	EndX	終了位置 X 座標	EndX	mm
13	EndY	終了位置 Y 座標	EndY	mm
14	PosZ	試料高さ	PosZ	
15	PointX	X 方向ポイント数	PointX	
16	PointY	Y 方向ポイント数	PointY	
17	Step	ステップ数	Step	
18				
19	['ResMap_Element']			
20	CompName	測定サンプル名	CompName	
21	ProcType	計測種類	ProcType	
22	Elem	観測元素種	Elem	
23	Line	観測 X 線名	Line	

データ利用者には，どのような利用ができるかがすぐにわかります．

DX化のValueは何か？

■ 機器利用者へ：　データ変換・データ成形などの時間のかかる「前処理」から解放する
■ データ利用者へ：　クリーニングでデータ構造化された「データセット」を提供する

現　状

機器利用の現状　　可視化まで１日がかり

実験装置A社

メーカーの解析ソフト
（兼バイナリー変換）

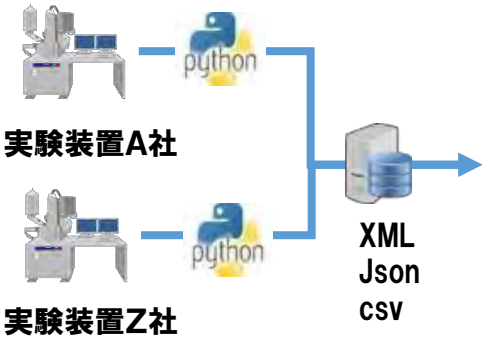
有償の科学解析ソフトで再解析

データ利用者の現状

- 特定のメーカのデータしか扱えない
- バイナリデータの解析がメーカー提供のコンバーターや解析ソフトに制限される
- 複数の解析ソフトを使って再解析する必要がある
- 最新のインフォマティックスのツールが提供されない、提供できない、コードが組めない

DX 化

機器利用のValue　　可視化まで１分以内



実験装置A社

実験装置Z社

XML
Json
csv

OSS：　例としてPythonのエコシステム

Matthew Rocklin (NVIDIA)　“Dask: Scaling Python”
https://goo.gl/LXUESM

データ利用者のValue

- データセットで一括したデータ集が得られる
- 複数のメーカのデータが統合できる
- 解析はOSSのエコスステムが利用できる
- 最新のインフォマティックスのツールで手軽に実行できる

例）
機械学習：　scikit-learn
画像解析：　Keras, TensorFlow

ご清聴ありがとうございました。

より詳しいことは

IoTデータ収集システムのデータアーキテクチャ
松波成行，松田朝彦，知京豊裕，原田善之，吉川英樹

情報処理学会トランザクションデジタルプラクティス　Vol.2 No.2
https://www.ipsj.or.jp/dp/contents/publication/46/TRO202-08.html