

NICT 高度通信・放送研究開発委託研究課題 採択番号:22609
データ利活用等のデジタル化の推進による社会課題・地域課題解決のための実証型研究開発

実世界データ醸造基盤の実現に向けて

NICT 第15回データ分析・可視化TF

株式会社ExData 代表取締役 永田 吉輝



株式会社ExData: 会社概要

- 名称: 株式会社ExData (エクスデータ)
- 設立日: 2022年7月7日 (設立から1年4か月)
- 社員数(非正規含): 8名
- 企業理念: データの力で、世界にワクワクを
 - 実世界のデータ収集・交換・分析を容易にし、誰もがデータから簡単に様々な知見や恩恵を得られるような世界を実現する
- 主な事業内容: データ分析など各種ソフトウェア開発

株式会社ExData: 代表取締役プロフィール

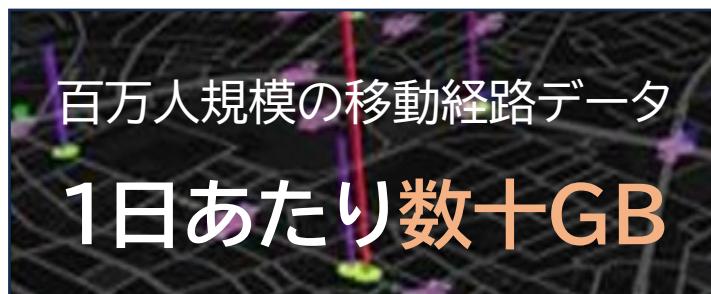
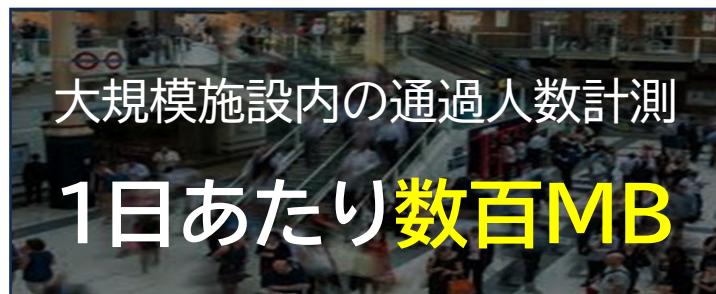
永田 吉輝 / Yoshiteru Nagata

所属

- 名古屋大学大学院 工学研究科 情報・通信工学専攻 D1
- 卓越大学院 ライフスタイル革命のための超学際移動イノベーション人材養成学位プログラム(TMI) 2期生
- 名古屋大学 融合フロンティアフェロー 量子科学分野 3期生
- 株式会社ExData 代表取締役

背景：実世界データ利活用の現状

交通・物流・気象・インフラなどのあらゆる分野で
大量の実世界データ収集による「**データ爆発**」が発生



実世界データの多くは、活用されずに**保管**または**処分**されている

何らかの価値を生み出すかもしれない
保管されているデータ

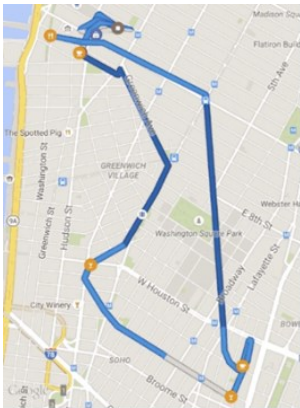
コスト・労力が課題となり...

価値を生み出すことなく
処分されたデータ

様々な人の移動データの例と特徴

移動データだけでも、様々な種類・特徴がある

人の移動データの種類例		データの特徴
主な管理者	データ	
情報通信事業者、 オンラインサービス事業者、 移動通信機器の管理者	衛星測位データ(GPS、準天頂衛星等)	アプリ利用者の正確な滞留・移動を把握可能。地下・室内では使いにくい。
	携帯電話・スマホの接続基地局データ	利用者の分布、移動を継続的に把握可能。精度はGPSデータに劣る。
	公衆無線LAN等の接続データ	電波の届く範囲での正確な滞留・移動状況を把握可能。地下・室内もOK。
	Bluetooth機器の接続データ	機種により、特定事項に関心がある人の来訪実態がわかる。地下・室内もOK。
	家庭用機器の利用データ	利用者の在宅時刻・時間、生活時間の実態把握が可能。データ量は増加見込み。
交通機関・道路管理者	交通施設(駅、バス停など)の通過データ	ほぼ悉皆に近い乗降客数を把握可能だが、時間と空間には制限あり。
	交通量データ(VICSのデータ)	道路・地域・広域の混雑状況の推移が把握可能。人の移動の正確な把握は困難。
店舗・ビル・公共施設等管理者、 決済サービス事業者	定点カメラの撮影データ	特定地点の混雑、人の滞留状況をリアルタイムで把握可能だが、夜間は困難。
	ビーコン等のセンシングデータ	定点カメラデータと同じようにリアルタイム性は高いが、人の属性はわからない。
	決済端末の決済データ	施設毎の消費行動の実態がわかるが、広域での移動状況は把握困難。
	入退館データIC	IDカードなどで管理している事業所・ビル毎の移動・滞留が把握できる。
公的機関	パーソントリップ・データ	統計的な精度を確保したデータ。マルチモーダルな移動や移動目的を把握可能。
	住民登録データ	ほぼ全国民の長期間の移動実態がわかるが、短期間の実態把握には不向き。
	高速道路通過、交通違反取締りデータ	ETC通過履歴やカメラから車ごとの移動移動情報を得られる。
	出入国管理データ	出入国履歴(法務省)と滞在地登録情報(外務省)から国外の移動情報がわかる。



経路



通過



滞留

数多くの種類・量の実世界データが“爆発”している

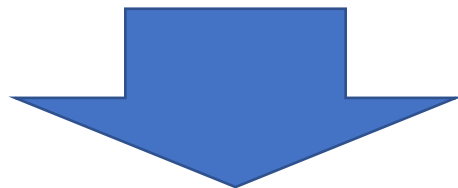
データ活用の実態

- ネットワークの整備やIoT機器の普及が進み、様々な実世界データを簡単に収集できるようになった
 - 人流・交通流データ
 - 購買データ
 - 気象データ
 - 金融データ
- 多くの組織では実世界データを、**持続的に活用できていない**
 - コロナウィルス感染症の流行以前のデータを既に削除してしまっており、コロナ以前との実世界データ比較を行えない
 - 様々なセンサでデータ収集を行い、データ分析を行おうとすると、各センサの仕様に沿ってプログラムを準備する必要があり、分析に非常に手間がかかる

このような社会の状態を、**データ・サステナビリティがない状態**と呼ぶ

データ・サステナビリティのために解決すべき課題

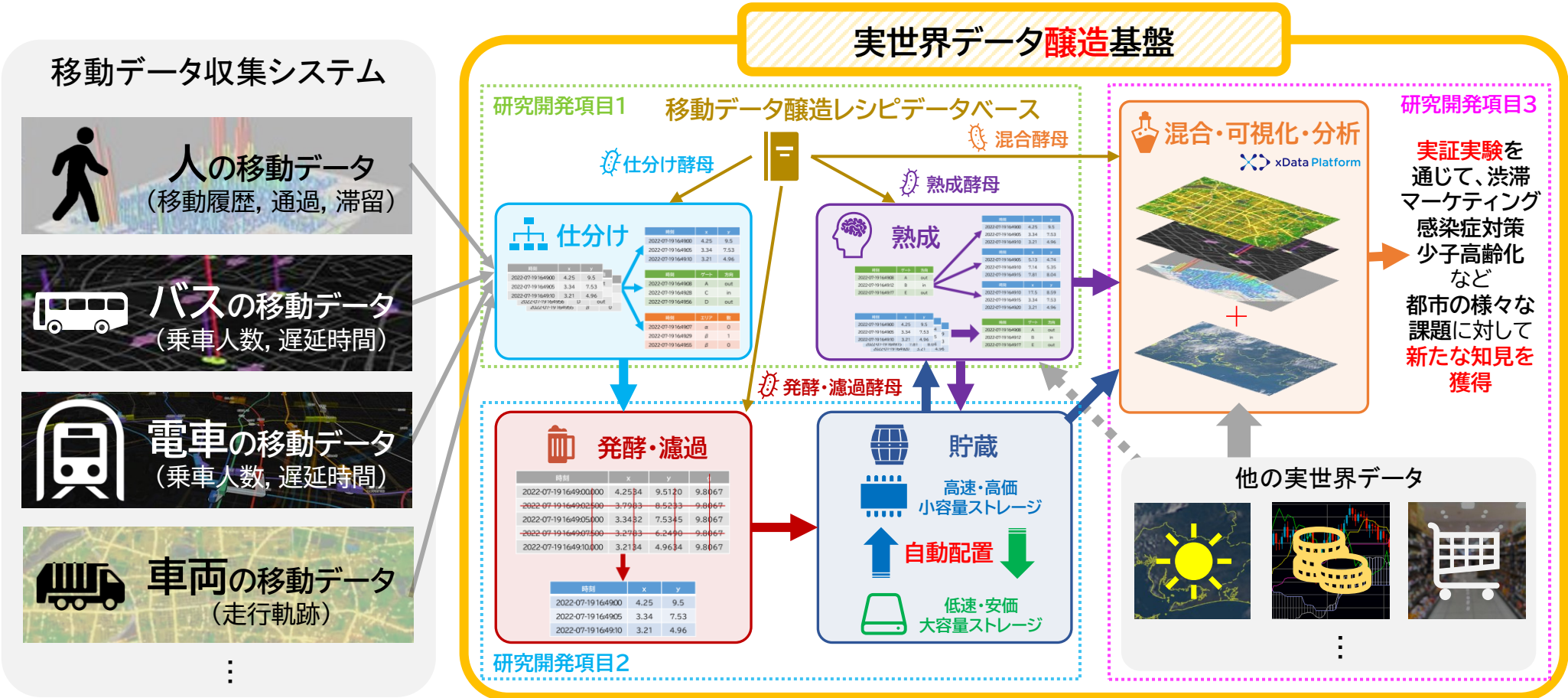
- データ収集: 異種実世界データを使いやすく整理した上で、横断的な分析を可能にする
- データ保管: 実世界データを効率的に圧縮し、有益なデータのみ保存する



実世界データ醸造により移動データの有効活用を可能にし、
データ・サステナビリティの実現を目指す

研究開発の概要

実世界データをサステナブルに活用するための基盤を開発



データ醸造とは？

一連のデータ利活用工程を酒類の製造工程に見立てたメタファ

- 一般的な酒類の醸造



- データ利活用の工程



「醸造」に含まれる技術要素を、データ利活用にも適用

- マイクロプログラム(=酵母)を複数用いたデータ処理
- 発酵・濾過と継続的な熟成による、有益なデータの抽出とムダの削減
- データを混合させ、新たな知見(=テイスト,味)の発掘を目指す

データ醸造による実世界データ利活用の将来

データ醸造なし

- 様々なIoTセンサからデータを個別に収集し、ステータス管理
- 収集したデータをストレージやデータベースに(無造作に)保存
- データに付随する様々なメタデータはデータ自体とは別で管理

- 様々なデータを保存時の形式のままひたすら蓄積
- 古くなったデータは圧縮してアーカイブするか、削除
- 使いたいときにデータがどこにあるかわからない…

- 様々なデータに対して専用の可視化・分析プログラムを実装
- 大量のデータを処理するため、可視化・分析処理にかかる時間が長い
- そもそもデータ分析のノウハウがなくてできない…

収集

保管

可視化・分析

データ醸造あり

- IoTセンサからの**データ収集状況を一元管理**
- データ保存時に、**保存先情報を醸造基盤で管理**
- データ**構造を自動認識**し、メタデータと共に管理

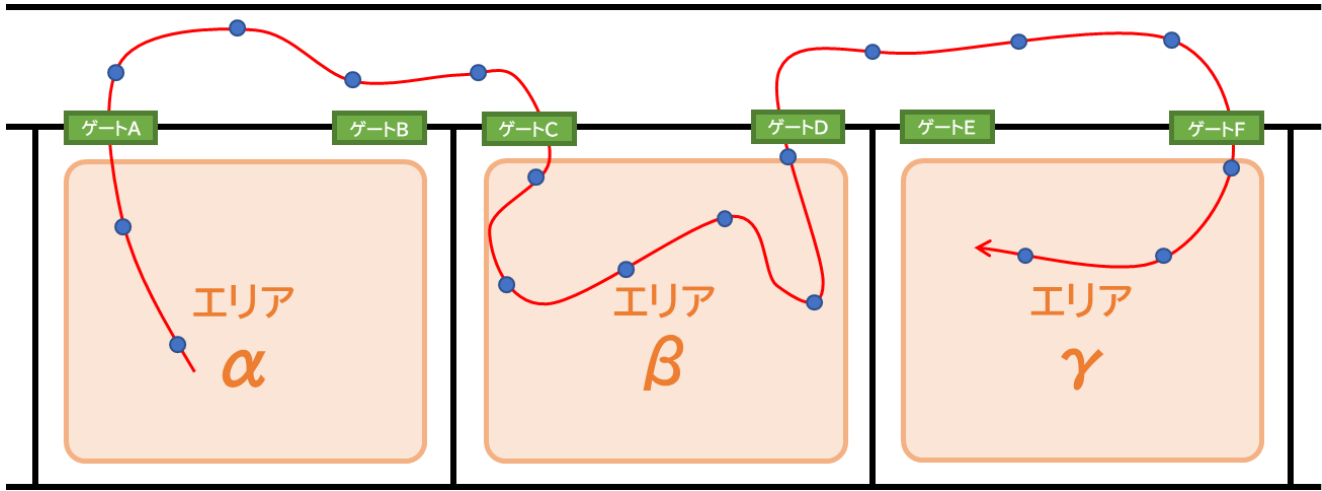
- 様々なデータに適した**データ醸造**を継続的に行い、よりデータ分析に適した**価値あるデータのみを保存**
- データの使用傾向を分析し、実際に使われているデータのみを高速なストレージに優先配置

- 様々な質的特徴を持つデータに対する可視化・分析プログラムを用意し、**同様のデータには共通のプログラム**を利用
- 醸造済みのデータを利用し、**高速な可視化・分析**を実現

様々な移動データ

単に移動データといっても、様々な種類が存在する

- 移動した**経路**を定期的に絶対座標で記録
- ゲートを**通過**した際にそのゲートの名前と通過方向を記録
- エリア内に**滞留**している人数の変化を記録



様々な
移動データ
が存在

経路

時刻	x	y
2022-07-19 16:49:00	4.25	9.5
2022-07-19 16:49:05	3.34	7.53
2022-07-19 16:49:10	3.21	4.96

通過

時刻	ゲート	方向
2022-07-19 16:49:08	A	out
2022-07-19 16:49:28	C	in
2022-07-19 16:49:56	D	out

滞留

時刻	エリア	人数
2022-07-19 16:49:07	α	0
2022-07-19 16:49:29	β	1
2022-07-19 16:49:55	β	0

研究開発項目1: データの仕分け

基盤に収集したデータを、質的特徴に基づき仕分け

- 経路: x,y 座標は絶対座標を表す量的変数
- 通過: ゲートはある地点を表す質的変数, 方向も質的変数
- 滞留: エリアはある領域を表す質的変数, 人数は量的変数



時刻	x	y	量的変数 & 量的変数
2022-07-19 16:49:00	4.25	9.5	
2022-07-19 16:49:05	3.34	7.53	
2022-07-19 16:49:10	3.21	4.96	

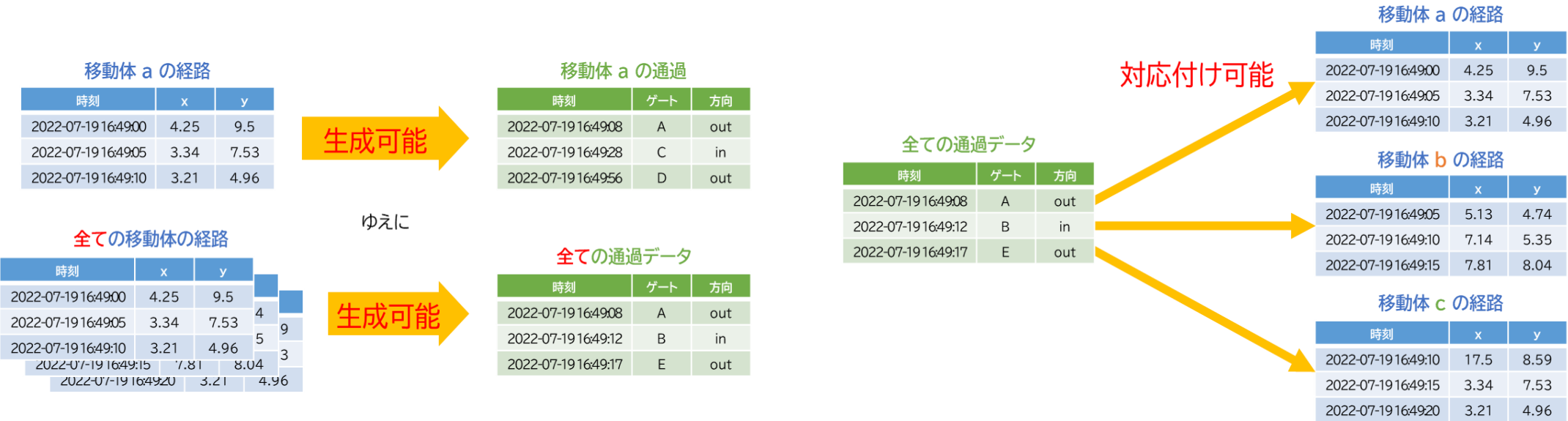
時刻	ゲート	方向	質的変数 & 質的変数
2022-07-19 16:49:08	A	out	
2022-07-19 16:49:28	C	in	
2022-07-19 16:49:56	D	out	

時刻	エリア	人数	質的変数 & 量的変数
2022-07-19 16:49:07	α	0	
2022-07-19 16:49:29	β	1	
2022-07-19 16:49:55	β	0	

研究開発項目1: データの熟成(継続的な変換)

データを継続的に・必要に応じて変換

- データを使う時期によって、必要な粒度・種類は変化
 - ex. 30年前のミリ秒/mm単位の個人の移動データ → 1時間にある地点を通った人数
- データ分析の内容によって、必要な形は変化
 - ex. ある地点を通った人数を知りたいが、経路データしか存在しない → 通過データへ



研究開発項目1: これまでの成果

課題: 質的・時空間粒度的に異なる移動データの仕分け・収集手法

- 様々な実世界データに対する統一的なメタデータ表現手法として、**schema.org** を拡張した **JSON-LD 形式でメタデータを表現**するスキーマを開発
- 計算機による自動的なデータ醸造を実現するため、実世界データの**構造情報をデータ自身から自動生成**し、実世界データ醸造基盤上で管理する枠組みを開発
- 様々な実世界データ収集システムのデータを、実世界データ醸造基盤上でその**収集状況や保管場所を管理**する枠組みを開発

JSON-LD形式によるデータ構造表現

実世界データのデータ構造を、
schema.org を拡張した
JSON-LD 形式で表現

- CSVやJSONなどの構造化データに対するデータ構造や変数の質的特徴を表現

→ データ醸造や複数の実世界データを交えた分析の自動化に必要なメタデータを生成

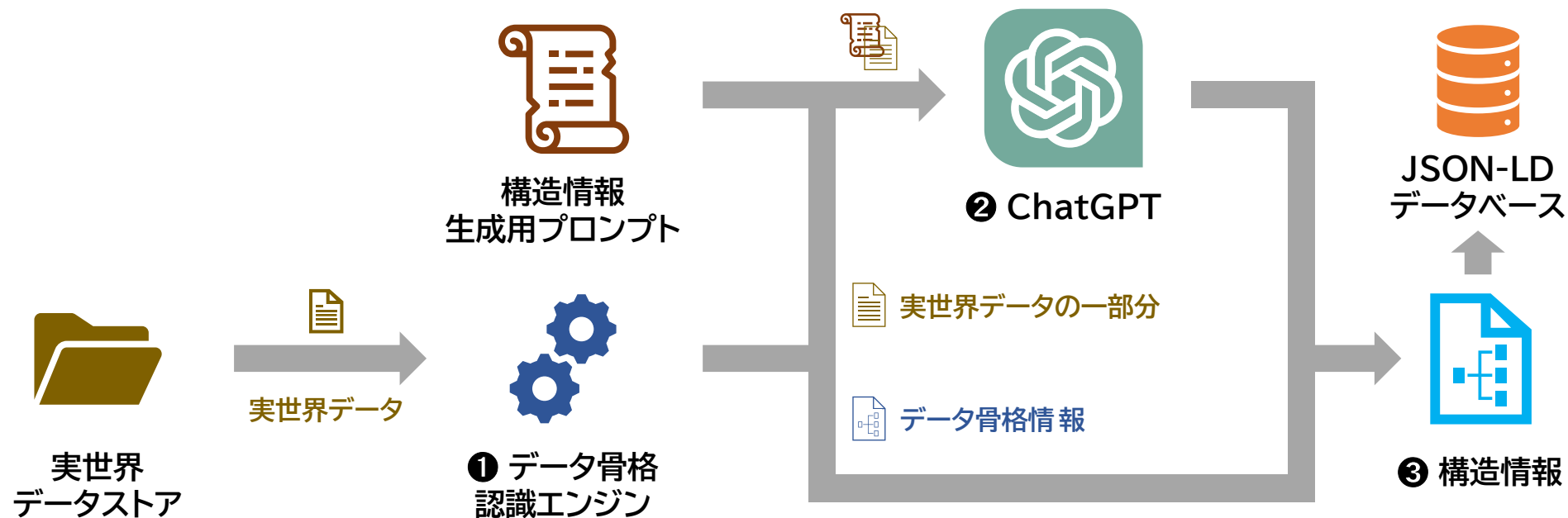
```
6909fcbb-ef24-4f9b-8981-4a81297cb133, 2023-01-01T00:00:00, 35.9879087, 136.883749, 178, 5986.023911  
ba3364b5-662e-4909-b6d1-0e59d8b34beb, 2023-01-01T00:01:02, 35.9870087, 136.883719, 168, 2346.051677  
72b81471-ed17-4f2d-8df7-1249f7daf08f, 2023-01-01T00:01:58, 35.9889087, 136.833749, 153, 6516.025145  
2159cb54-b22d-4fa8-8a12-7eeeb41f4348, 2023-01-01T00:03:00, 35.9679087, 136.883449, 181, 59516.021549  
98f1b873-31fc-4d6c-a3c6-3ca562830c14, 2023-01-01T00:04:01, 35.9179187, 136.853759, 175, 99516.021966
```



```
{  
  "jsonld": {  
    "@context": { ... },  
    "@type": "dbp:structureInfo",  
    "@graph": [{  
      "@id": "root", "@type": "rdf:List", "rdfs:label": "root",  
      "schema:rangeIncludes": {"@id": "class1"}  
    }, {  
      "@id": "class1", "@type": "rdfs:Class", "rdfs:label": "class1",  
      "schema:domainIncludes": {"@id": "root"},  
      "schema:rangeIncludes": [  
        {"@id": "column0"}, {"@id": "column4"}, {"@id": "column2"}, {"@id": "column5"}, ...  
      ]  
    }, {  
      "@id": "column0", "@type": "dbp:RealWorldDataStructureProperty",  
      "rdfs:label": "column0", "dbp:itemType": "string",  
      "schema:domainIncludes": {"@id": "class1"},  
      "schema:rangeIncludes": {"@id": "schema:Text"},  
      "dbp:VariableCharacteristicEnumeration": "名義尺度",  
      "rdfs:comment": "ランダムなUUID (Universally Unique Identifier) の値"  
    }, {  
      "@id": "column1", "@type": "dbp:RealWorldDataStructureProperty",  
      "rdfs:label": "column1", "dbp:itemType": "string",  
      "schema:domainIncludes": {"@id": "class1"},  
      "schema:rangeIncludes": {"@id": "schema:Text"},  
      "rdfs:comment": "時刻を表すデータ"  
    }, ...  
  ]  
}
```

構造情報の自動生成

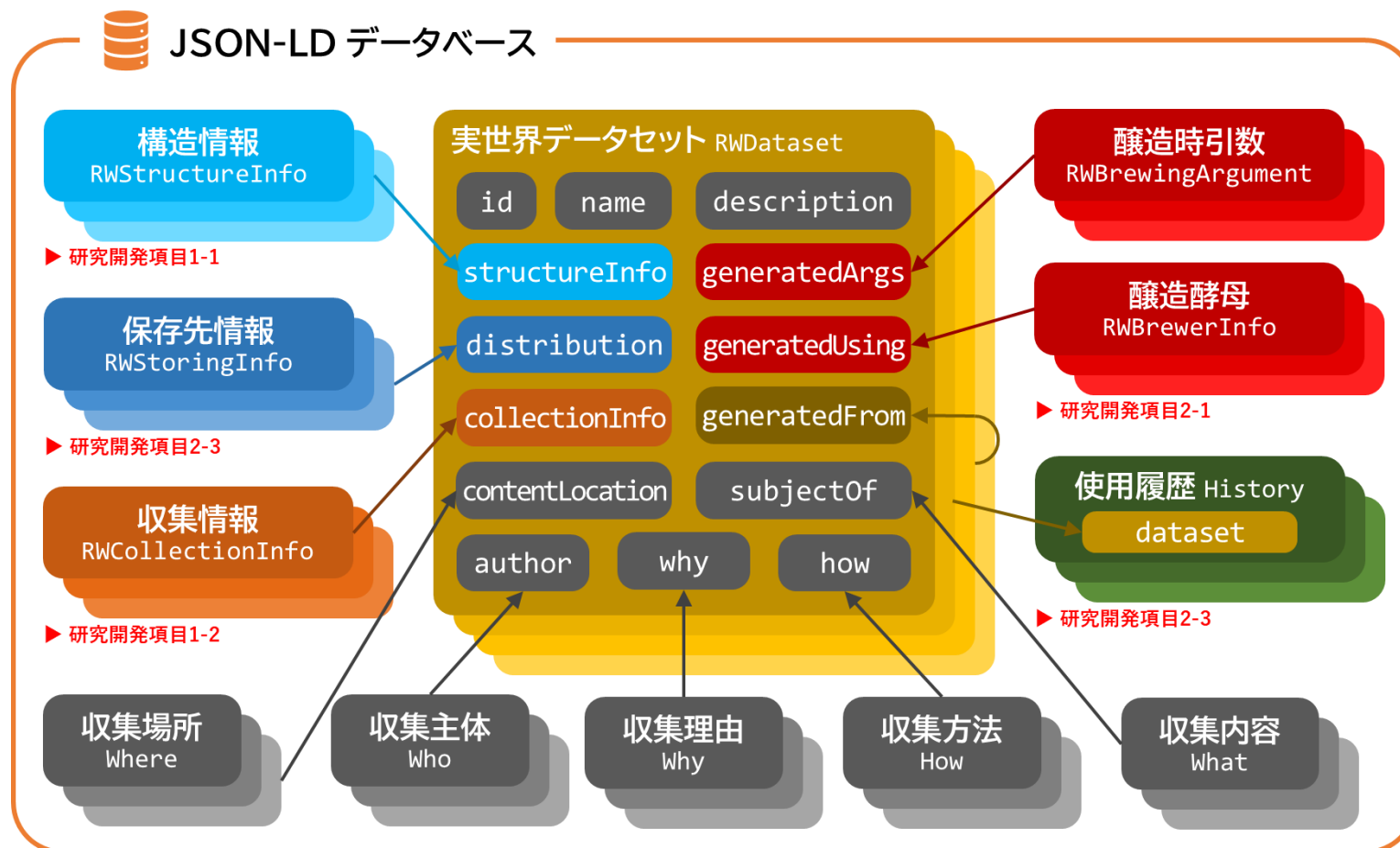
実世界データの構造情報を、生成AI技術を併用し自動生成



→ データ整理にかかる手間を削減 & Linked Data 化を推進

JSON-LD データベース

実世界データの様々なメタデータを JSON-LD データベースとして管理



JSON-LD 間のリンク

JSON-LD 間のリンクで詳細なメタデータを分散して管理

→ 同種の実世界データには共通のメタデータを再利用し、検索性などを向上

実世界データセットの JSON-LD

```
{
  "@context": { ... },
  "@type": "dbp:RWDataset",
  "name": "Example で観測した人の通過データ",
  "author": { "@id" : "... " },
  "contentLocation": { "@id" : "... " },
  "structureInfo":
    { "@id" : "https://example.com/data-structure" },
  "distribution": { "@id" : "... " },
  ...
}
```

▲ <https://example.com/datainfo>

<https://example.com/data-structure> ►

構造情報の JSON-LD

```
{
  "@context": "https://exdata.co.jp/dbp/schema/",
  "@type": "dbp:RWDataStructureInfo",
  "schema:name": "ExampleSensor の出力形式",
  "schema:encodingFormat": "text/csv",
  "dbp:structure": {
    "@type": "dbp:structureInfo",
    "@context": { ... },
    "@graph": [
      { ... },
      ...
    ],
    ...
  },
  ...
}
```

実世界データ収集時のメタデータ・ステータス更新

既存のデータ収集システムに容易に組み込める仕組みを開発

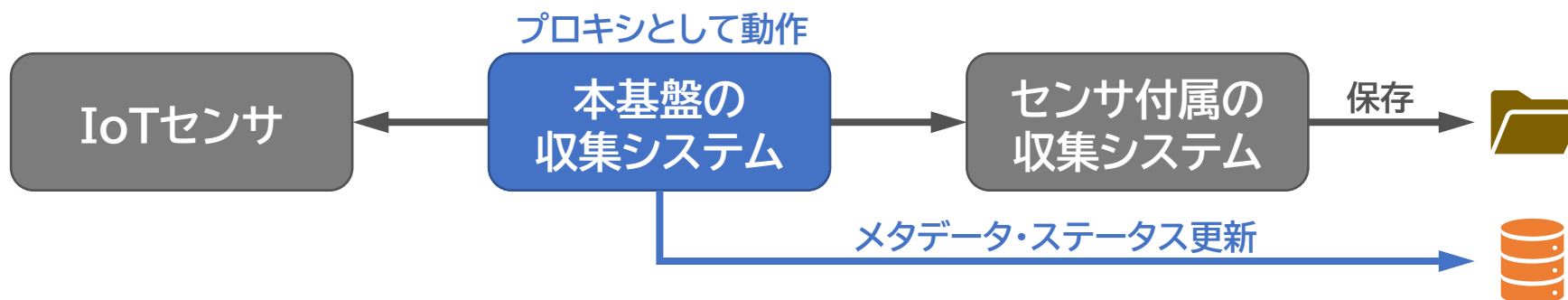
従来型の
データ収集



本基盤による
データ収集
(再現型)



本基盤による
データ収集
(プロキシ型)



JSON-LD データベース ウェブクライアント

The screenshot shows the JSON-LD Database Web Client interface. The left sidebar contains a menu with items: ダッシュボード, 可視化, 実世界データ, 実世界データセット, 構造情報, 収集情報, 保存先情報, データ醸造, 醸造プログラム, 定期醸造設定, 定期移動設定, 定期削除設定, 醸造入力, 醸造出力, and 醸造指数. The main content area has a header with '実世界データ醸造基盤' and a search bar. Below the header, there is a section for '実世界データセット' with a table of datasets. The table has columns for Name, ID (JSON-LD URL), and 収集情報. The table lists five datasets: サンプルセンサ, 倉庫録画データ_カメラ01, 中部国際空港_人流通過センサ, 気象庁データ, and 新規実世界データセット. A red dashed box highlights the search bar and the table. A red triangle points to the search bar with the text 'JSON-LD データの検索'. A red triangle points to the table with the text '個別 JSON-LD データの参照・編集'. A red arrow points to the sidebar with the text '各種 JSON-LD データを参照'.

実世界データ醸造基盤

Search...

SEARCH

▲ JSON-LD データの検索

実世界データセット + 実世界データセットを追加

☐ Name	ID (JSON-LD URL) ↑	収集情報
☐ サンプルセンサ	/api/v0/dataset/1/?format=json	/api/v0/collection_info
☐ 倉庫録画データ_カメラ01	/api/v0/dataset/17/?format=json	/api/v0/collection_info
☐ 中部国際空港_人流通過センサ	/api/v0/dataset/23/?format=json	/api/v0/collection_info
☐ 気象庁データ	/api/v0/dataset/41/?format=json	/api/v0/collection_info
☐ 新規実世界データセット	/api/v0/dataset/56/?format=json	/api/v0/collection_info

▲ 個別 JSON-LD データの参照・編集

圧縮ビュー

◀ 各種 JSON-LD データを参照

研究開発項目2: データの発酵・濾過(抽出・圧縮)

データから有益な部分を抽出し、ムダな部分を継続的に削除

- 例: 表形式の経路データの場合 (CSV, etc.)
 - 分析に必要なデータは**残す**(時刻, x, y)
 - 全行が同じデータや、分析とは無関係な情報の列は**削除**
 - 必要以上な桁数(精度)を持つデータは桁数を**削減**
 - 必要以上な頻度で記録されたデータは、一部**間引く**

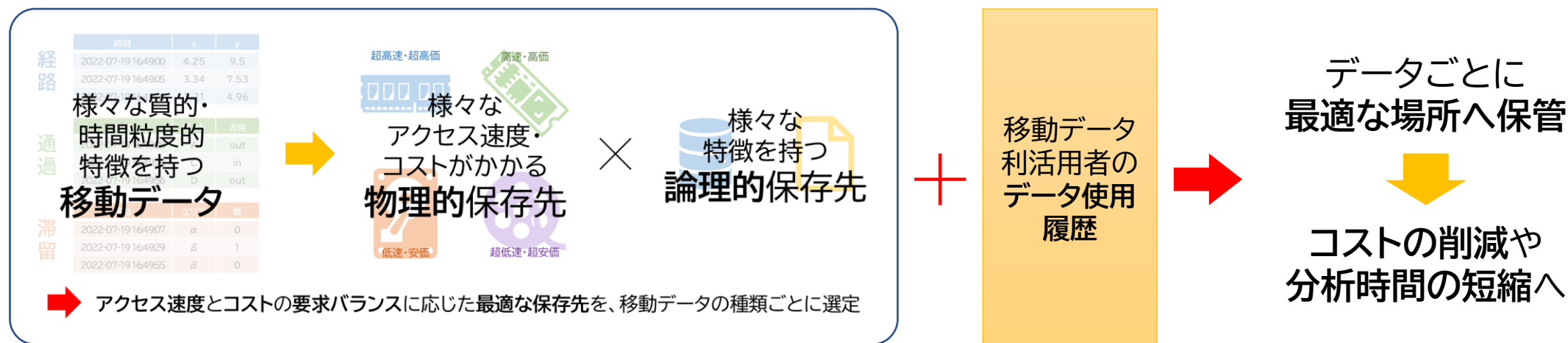
前スライドの
各分類に対し
最適な抽出・
圧縮法を定義



研究開発項目2: データの貯蔵

データの特徴や使われ方に応じて, コスパの良い保存先に自動貯蔵

- 様々な移動データに対し, コストと速度のバランスに応じ, 保存先の候補を事前に選定
- データ利活用者のアクセスパターンに基づき, データを自動配置・必要に応じて移動



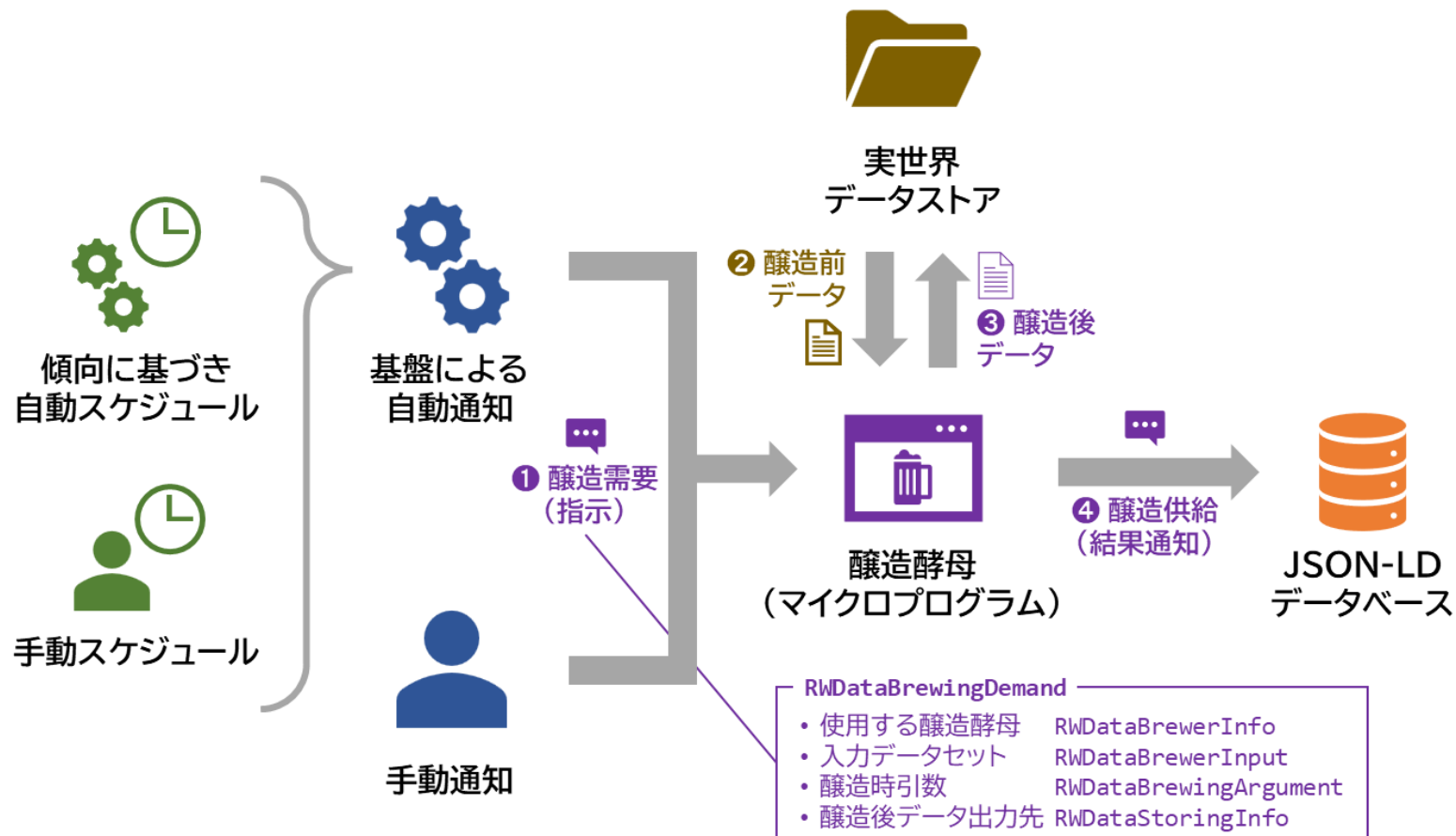
研究開発項目2: これまでの成果

課題: 効率的な移動データ発酵・濾過・貯蔵技術の研究開発

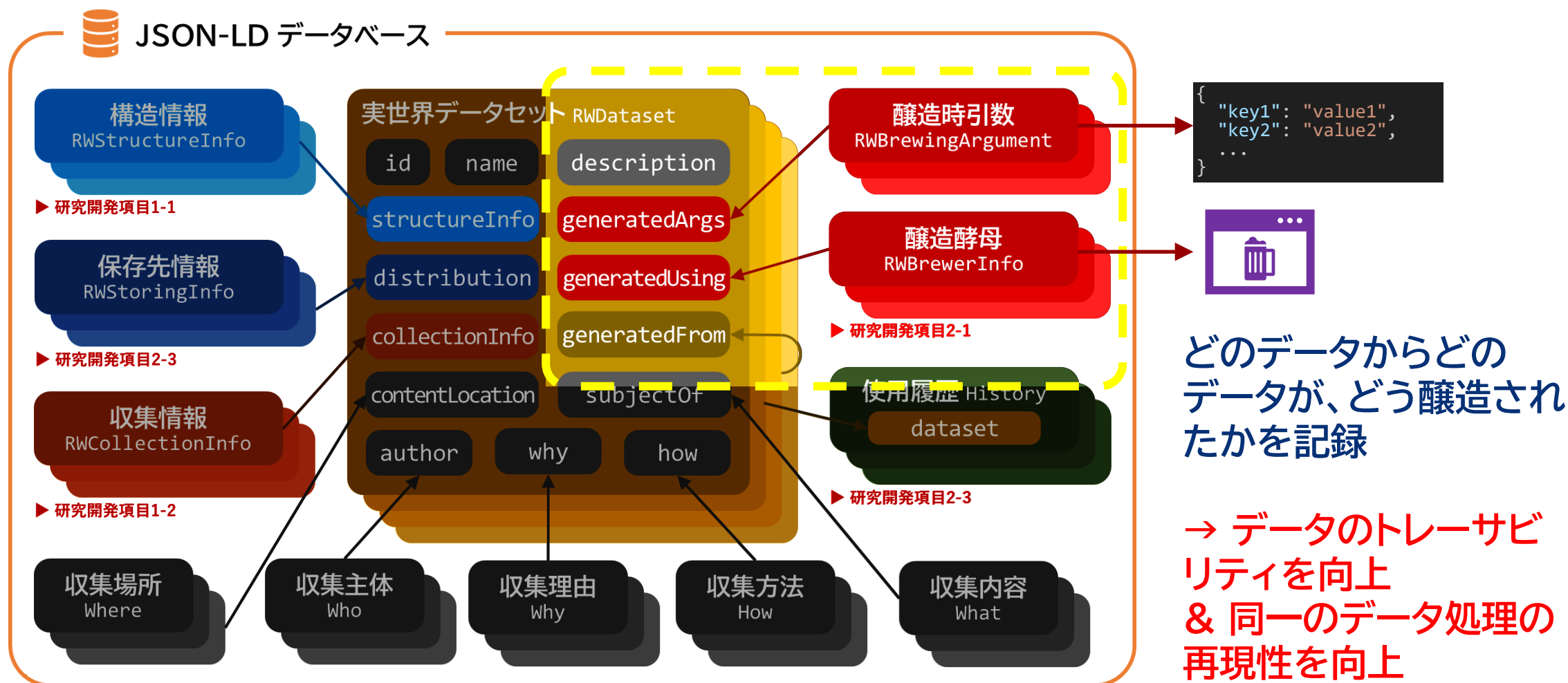
- 実世界データの**トレーサビリティ**を意識した、醸造酵母の基本設計と、酵母の機能を表現するメタデータのスキーマを開発
- 実世界データから価値ある部分を抽出(発酵)するための異常検知と、不必要な部分を削減(濾過)する数値処理、映像データの圧縮を行う**サンプル酵母を実装**
- 様々な種類のデータ貯蔵先の性能とコストを評価し、実世界データを**適切な貯蔵先に保存**する仕組みを実装
- 実世界データの**利用状況に応じて、自動的に発酵・濾過・貯蔵の醸造処理**を行うためのデータ基盤を実装

データ醸造の流れ

多様なデータに対応する汎用的なデータ醸造の流れを開発



データ醸造関連の JSON-LD データベース



サンプル醸造酵母

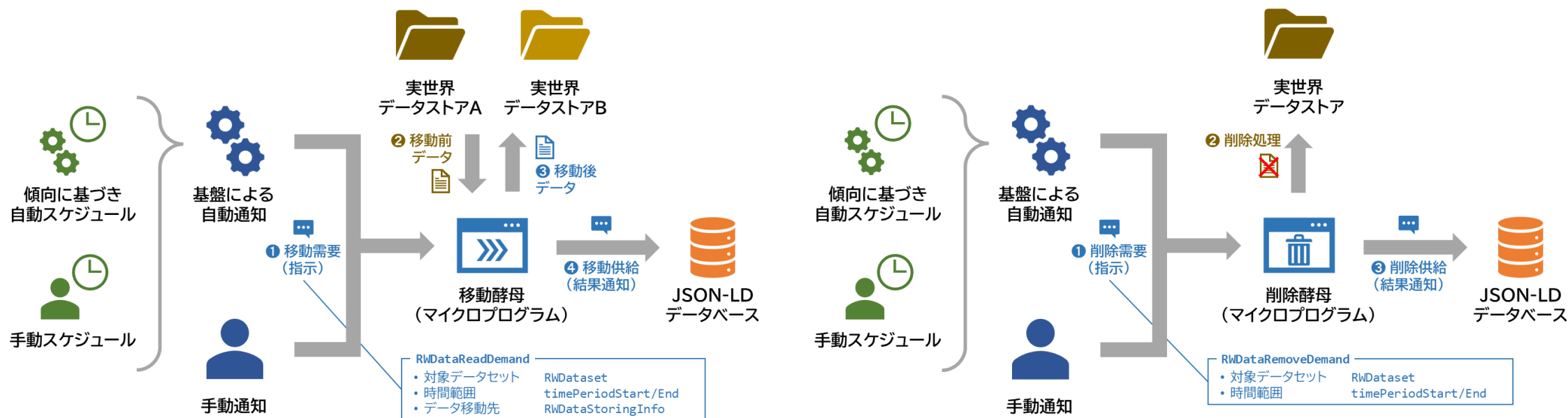
実世界データの発酵・濾過を行う3種類の酵母をサンプル実装

- 異常検知: 時系列データ中の特異な点を検出（発酵）→ データから**価値のある部分を抽出**
 - 入力: 構造化された時系列データセット
 - 引数: 処理対象とする構造部分（CSVのカラムやJSONの特定のキー等）
 - 出力: 時系列に沿った各データ点がどれだけ特異かを表すスコアを追加
- 数値処理: データ中の不必要・無意味な数値桁数を削減（濾過）→ データの**無駄を削減**
 - 入力: 構造化された時系列データセット
 - 引数: 処理対象とする構造部分、丸める桁数
 - 出力: 時系列に沿った各データ点がどれだけ特異かを表すスコアを追加
- 動画データの圧縮: コーデックやビットレートを変換し圧縮（濾過）→ データの**無駄を削減**
 - 入力: 動画データ
 - 引数: 変換後のコーデック、ビットレート
 - 出力: 時系列に沿った各データ点がどれだけ特異かを表すスコアを追加

貯蔵したデータの移動・削除

醸造時と同様の流れで、データの移動・削除を行う

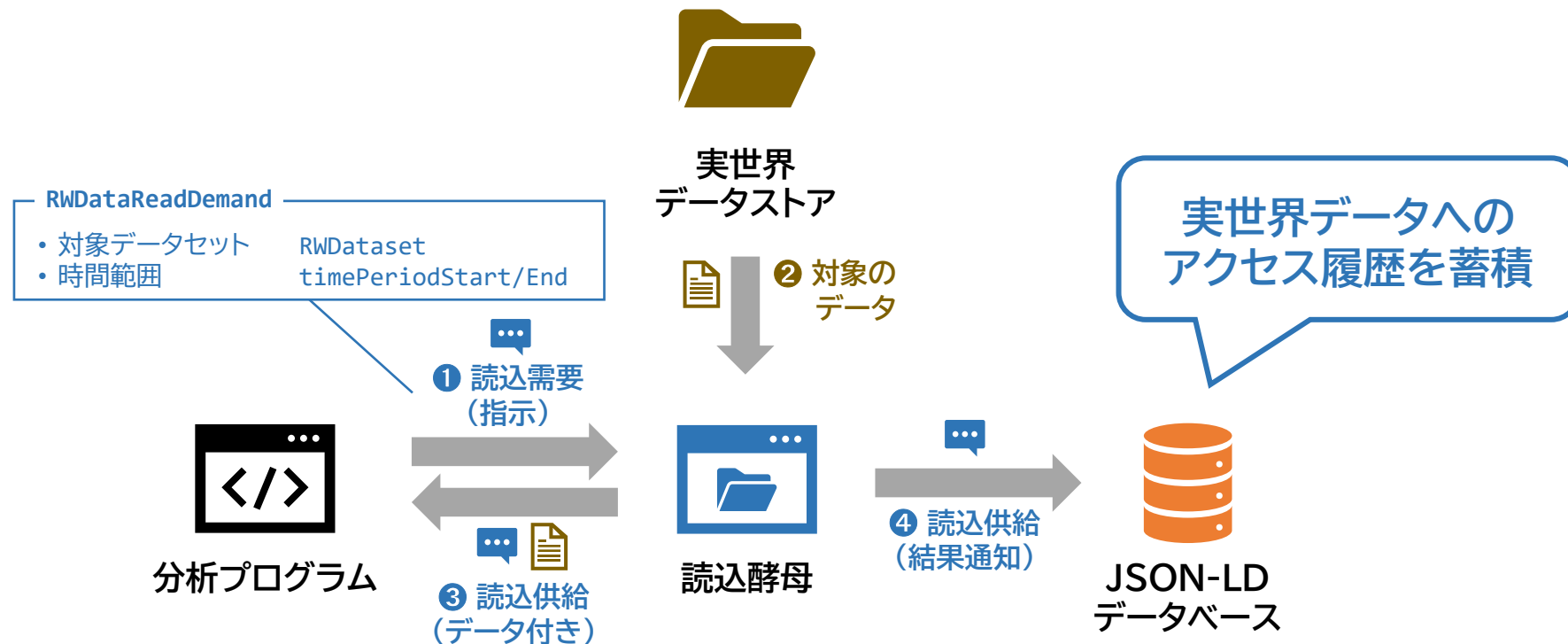
→ 移動後のデータの所在や削除した履歴を JSON-LD データベースで把握



実世界データへのアクセス履歴の収集

多彩な実世界データに対して統一的なデータ読込プロトコルを定義

→ データアクセスを簡素化しつつ, 需要に基づく自動配置のためのデータを蓄積



研究開発項目3：データの混合・可視化・分析

複数の実世界データから、新たな社会課題解決を実現

本基盤で醸造した移動データ

時刻	x	y	
2022-07-19 16:49:00	4.25	9.5	同
2022-07-19 16:49:05	3.34	7.53	t
2022-07-19 16:49:10	3.21	4.96	
2022-07-19 16:49:56	D	out	
2022-07-19 16:49:55	B	0	



実世界移動データと他のデータの
相関や潜在表現を抽出

混合

他の実世界データ



Harmoware-VIS

移動データと時空間データを
重ね合わせて可視化

都市の様々な
課題に対して
新たな知見を
獲得

研究開発項目3： これまでの成果

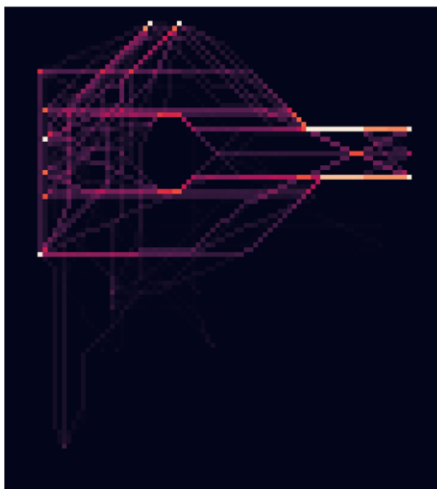
課題：異種実世界データの可視化・分析技術の研究開発

- 通過・滞留・経路の3種の移動データを可視化するシステムを開発
- 人流の予測モデルにデータ醸造前後の移動データと交通データを使用し、**人流予測精度**に与える影響を評価
- 動画データを圧縮する醸造技術が、倉庫内映像データの**文字認識やインスタンスセグメンテーション**に与える影響を評価
- NICT xData Platform 上で、気象データと移動データを連携して分析するサンプルプログラムの実装を開始

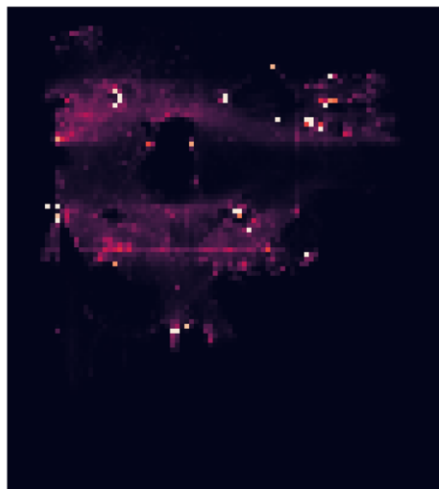
移動データの様々な可視化システム

移動データとその分析結果の様々な可視化システムを開発

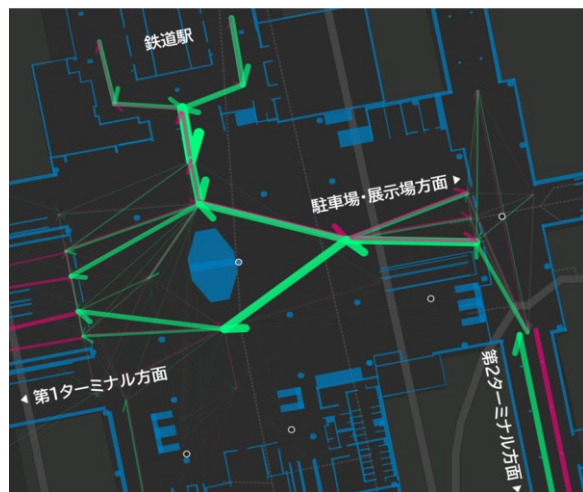
推定経路に基づく滞留予測



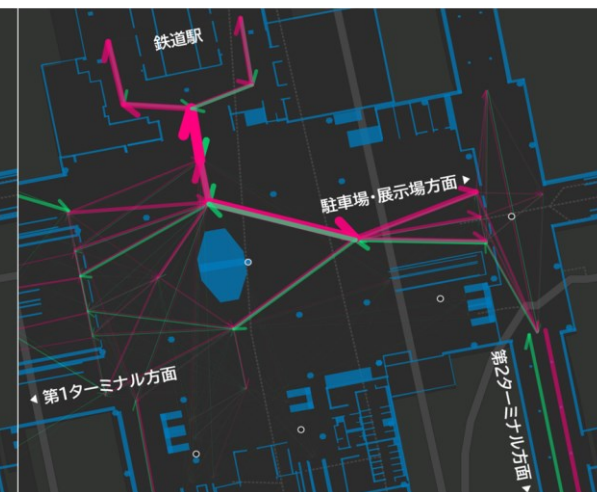
追跡データによる滞留



午前



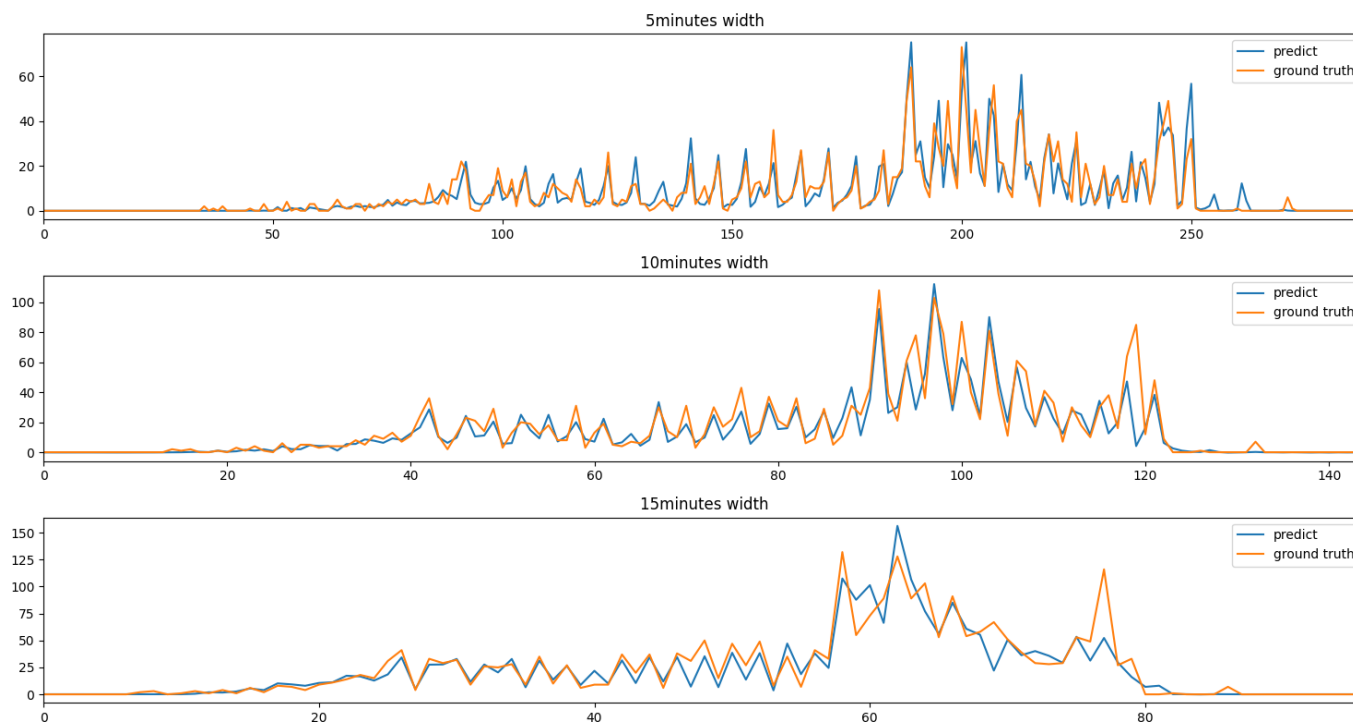
午後



→ 研究開発項目1 で開発した JSON-LD メタデータに基づき、
各実世界データに対する**適切な可視化を自動化**する仕組みを開発中

人流予測精度への影響評価

- 人の通過数と鉄道の発着状況から、将来の人流予測を行う
 - 学習・推論時に、5分 or 10分 or 15分単位の人々の通過数を使用
 - 必要な時間間隔に応じて醸造度合いを調整可能だと判明



倉庫内映像の文字認識・インスタンスセグメンテーション



倉庫内映像認識におけるデータ醸造の影響

- 元動画の h.264/3120kbps に対して、1/4 のサイズとなる h.265/780kbps までは、影響は認められなかった
 - 人目で見ても、多少の残像感はあるけどほとんど見分けがつかないレベル

圧縮方式/ビットレート	タイムスタンプ認識	物体認識
h.265/1560kbps	正常に認識	影響なし
h.265/780kbps	正常に認識	影響なし
h.265/390kbps	正常に認識	映像が乱れるとやや影響あり
h.265/195kbps	少数の誤認識が発生	映像が乱れると影響が大きい
h.265/98kbps	多数の誤認識が発生	認識困難
AV1/1560kbps	-	影響なし
AV1/780kbps	-	映像が乱れると影響あり
AV1/390kbps	-	映像が乱れると影響が大きい
AV1/195kbps	-	映像が乱れると影響が大きい
AV1/98kbps	-	認識困難

まとめ

様々な実世界データを扱った経験から、
大規模データを持続的に保持できない、という課題を再認識

⇒組織によっては数カ月から数年で貴重なデータを処分

持続的にデータを利活用する「**データ・サステナビリティ**」を提唱

「**醸造**」のメタファにより、ワインを貯蔵・熟成させるように、仕分け・発酵・
濾過・貯蔵・熟成といった継続的・長期的な実世界データ処理を実装

⇒データを実社会で持続的に活用する「**実世界データ醸造基盤**」を構築