# データジャケットを用いたデータ利活用方法・データ市場創出

東京大学大学院工学系研究科

システム創成学専攻 大澤研究室

早矢仕 晃章

# データ利活用に対する期待の高まり

## Big Data

- 保存可能なデータ量が増加
- 多様なデータに対応した分析手法の提案

## 個人情報端末の普及

- パーソナルデータが取得可能に
- 消費者の購買履歴、移動記録、興味など

## オープンデータ

- 二次利用可能なデータの公開
- 膨大な行政のデータにアクセス可能に

## センサーの高度化

- Internet of Thing、ドローンなど
- 高粒度のデータ取得が可能に

組織を越えた異なる領域のデータを組合せて新たな知識を獲得し、意思決定に役立てることへの期待が高まる（McKinsey Global Institute 2013, 2015など）

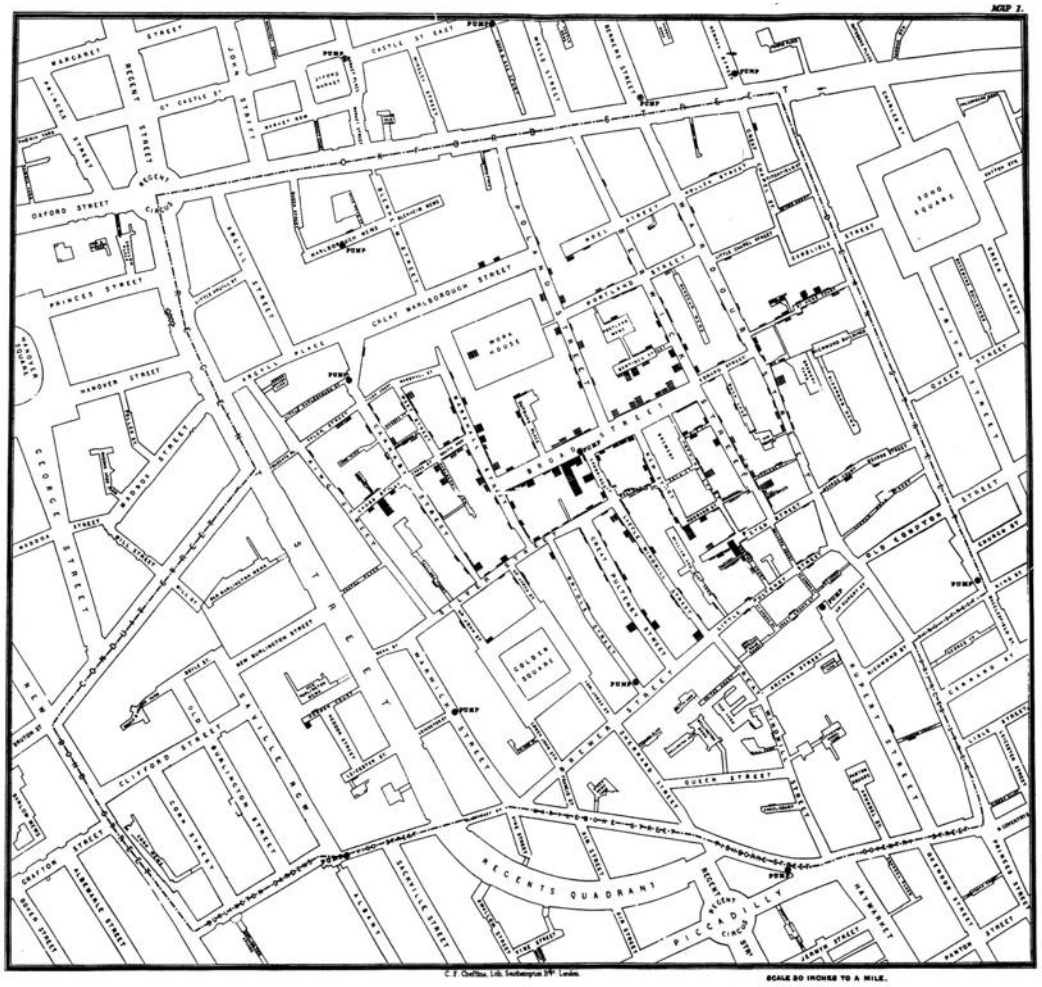
→既存ビジネスの付加価値向上

→新規事業の開発

# The world's most valuable resource is no longer oil, but data


The Economist誌（2017年5月）

# コレラの死者と井戸の位置データの組合せ（John Snow）




https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak#/media/File:Snow-cholera-map-1.jpg

# データ市場の萌芽

Webをプラットフォームとし、新たなデータ市場が形成

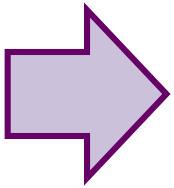






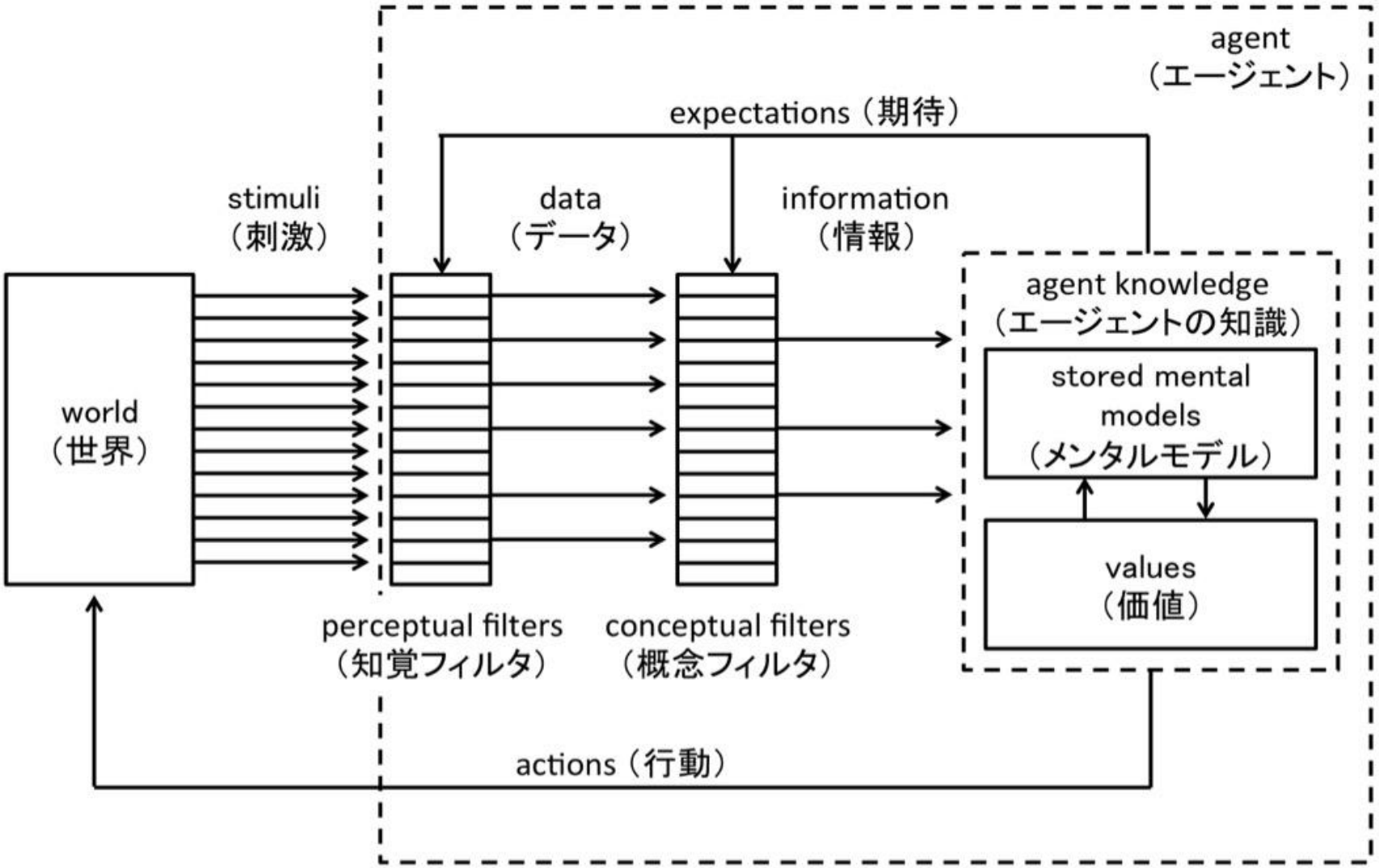



そのほか多数



データ流通をキーワードに、プラットフォーム間のデータ連携、異分野データ連携に対するニーズとカタログ整備の必要性が増してきている。

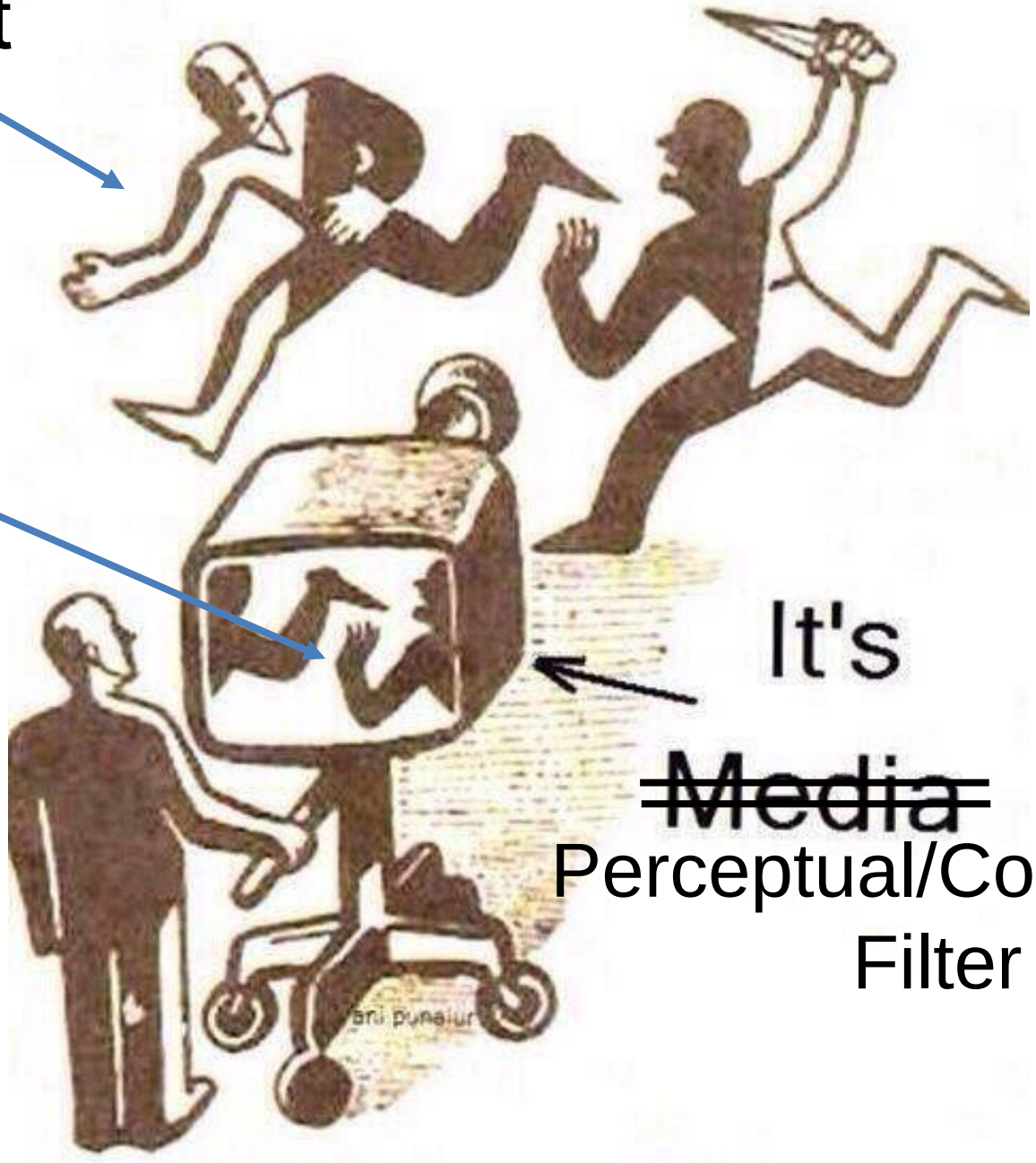
# データ・情報・知識




The agent-in-the-world [出典：Boisot & Canals, 2004. 筆者一部改変]

It’s Event

It’s Data



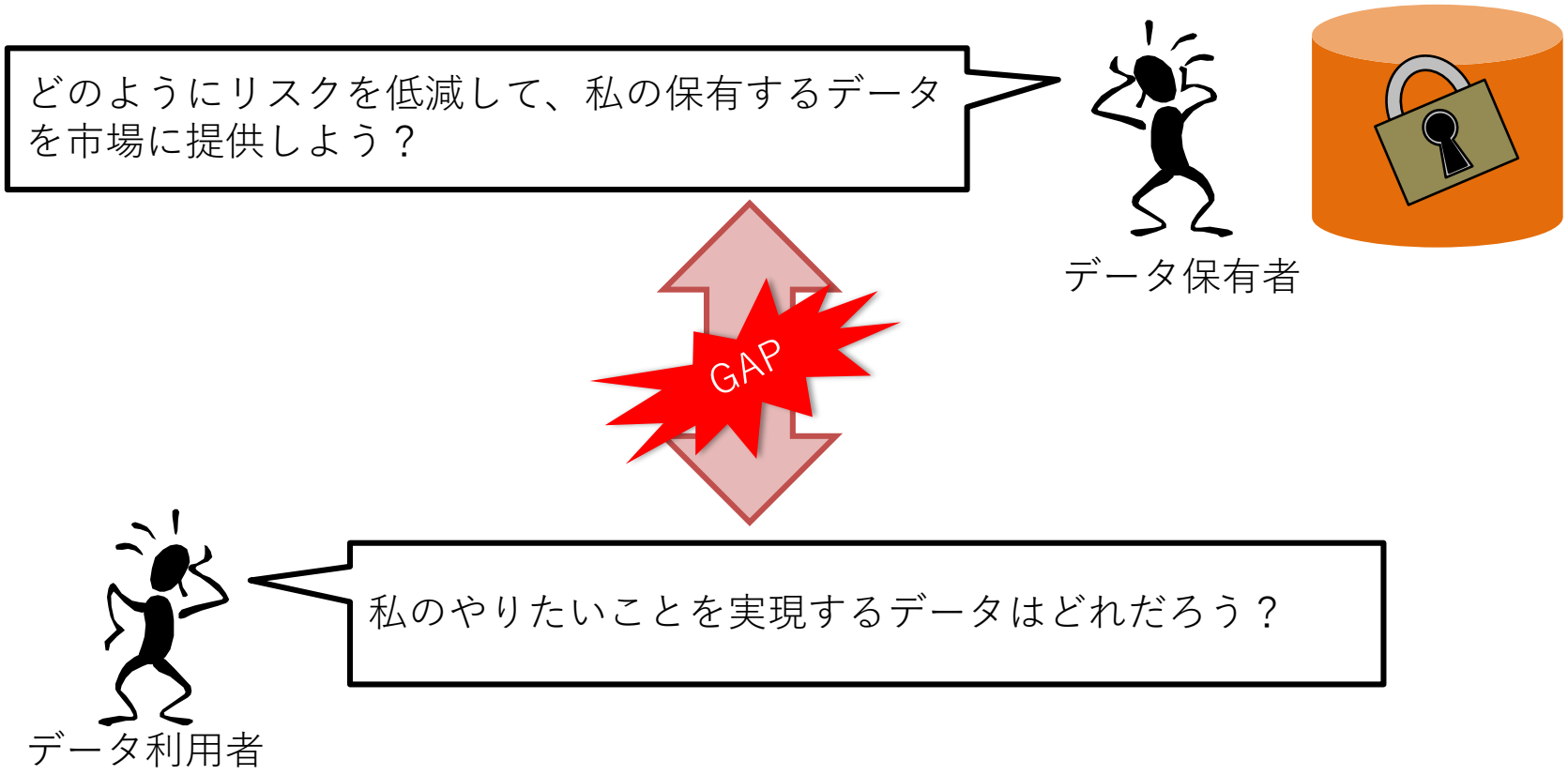
It's

~~Media~~

Perceptual/Conceptual Filter

# データ保有者と利用者の間のギャップの解決

- 個人、企業、その他の研究機関で収集・蓄積されているデータの存在を知るための手段はほとんどない。
- プライバシー侵害やビジネスの機会損失のリスクを低減してデータ保有者が持つデータを市場に提供する方法は確立していない。



# 解決すべき問題

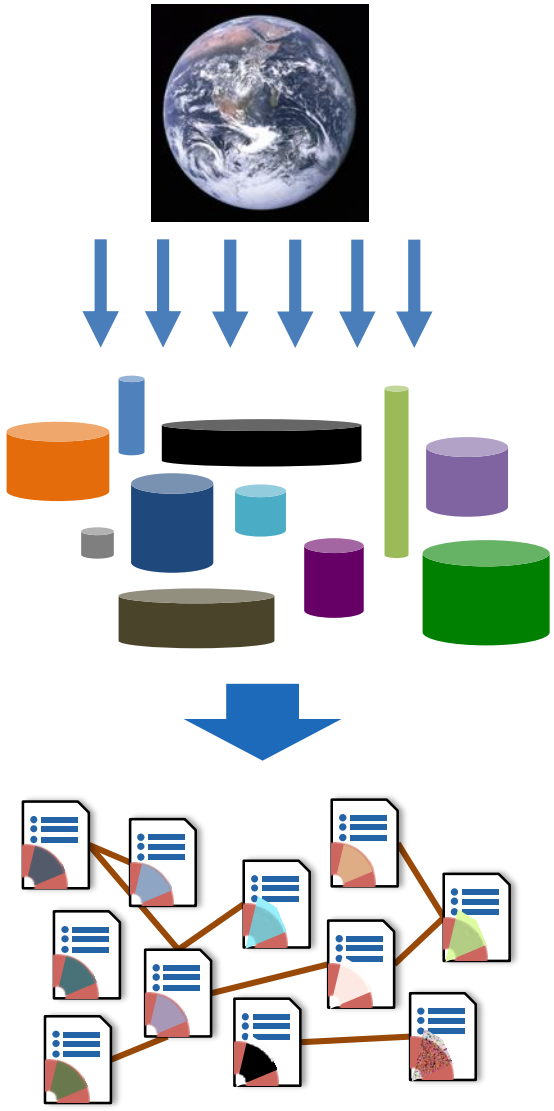
**異分野データ連携によるイノベーションを実現するためには・・・**

世界の見方、世の中に存在するデータの構造とそれらの関係を正しく理解することが重要

個々のデータの分析ではなく、データによって構成されるデータの母集団がどのような**構造的特徴**を有しているのかを調べること

データ全体の傾向や特徴を定量的に評価するためには、多様なデータを同じ土俵で議論するためのモデルが必要

データのデータ、すなわちメタデータを分析対象として利用することが有効



# データジャケットData Jacket

（Ohsawa et al., 2013）

- データジャケット（DJ）は人間が読むことを前提としたデータの概要情報
- 自然言語で記述され、構造化されている
- データの中身ではなく、概要情報（変数名、保存形式、収集方法など）を共有し、データの利用価値を検討可能にする
- 個人情報を含む共有不可能なデータでも、DJにすることでセキュリティ上のリスクを低減させて情報が共有可能となる
- 構造化によって、人間だけでなく計算機においても可読化

| 年 | 月 | 日 | 顧客ID | 購入品目 | 支払金額 |
|---|---|---|---|---|---|
| 2017 | 11 | 1 | AAAAA | 人工知能学会誌 | 2592 |
| 2017 | 11 | 1 | BBBBB | ペン、りんご、パイナップル | 1080 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2017 | 11 | 30 | YYYYY | スナック、するめ、ビール | 2536 |
| 2017 | 11 | 30 | ZZZZZ | ラーメン、ナタデココゼリー | 867 |

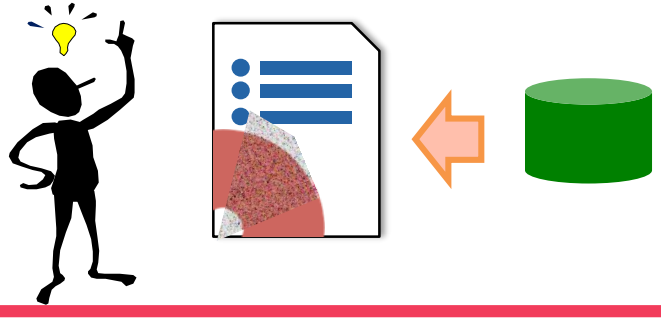
DJとして記述

| DJ No. XX【購買履歴データ】 | |
|---|---|
| 概要 | 東京都の◯◯スーパーマーケットで収集されている顧客の購買行動履歴。 |
| 収集方法・コスト | ポイントカードとPOSによって取得 |
| 共有条件 | 共有不可 |
| データの種類 | 表形式、テキスト、数値 |
| 保存形式 | CSV |
| 分析・シミュレーション | 時系列分析 |
| 変数ラベル | 氏名、性別、顧客ID、支払金額、購入品目、日にち |
| 分析結果 | • その日の売上の計算<br>• 今後の売上の予測と仕入れの推定 |
| 期待される分析 | 顧客の購買行動とリピート率を計算し、ロイヤルカスタマーの特定が可能かもしれない。 |
| コメント | 有効なデータの組み合わせが発見されればデータの提供あるいはコラボレーションもあり得る。 |

# データジャケットで出来ること

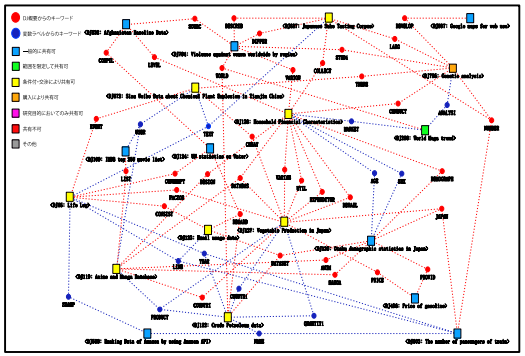
- 誰が・どこに・どのようなデータを・どのような形で保有しているのか理解することが可能となる。
- 異なるフォーマットのデータをメタデータとして記述することで統一的に扱えるようになる。

データ理解



データ検索

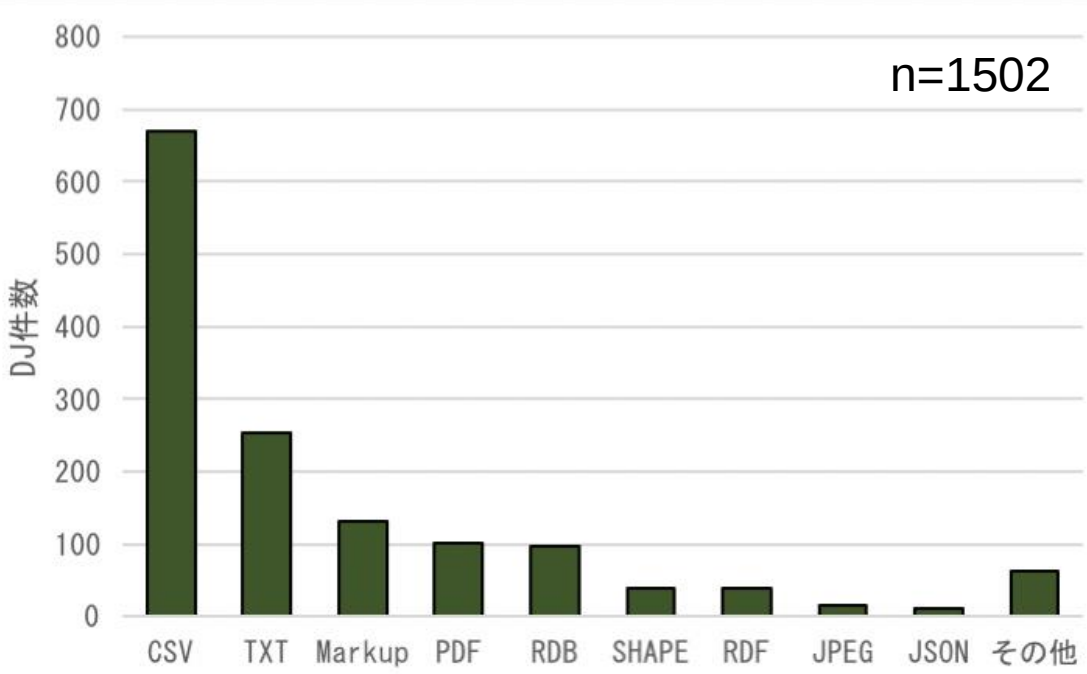
データ可視化



活用方法の検討

# データ理解

# DJから分かるデータの属性と特徴




CSV

国別のGDPデータ、行政のデータ

TXT

癌患者へのインタビューデータ

Markup

日本の火山データベース

PDF

授業における指導案と教材

RDB

主要地点の交通量

SHAPE

流域下水道データ

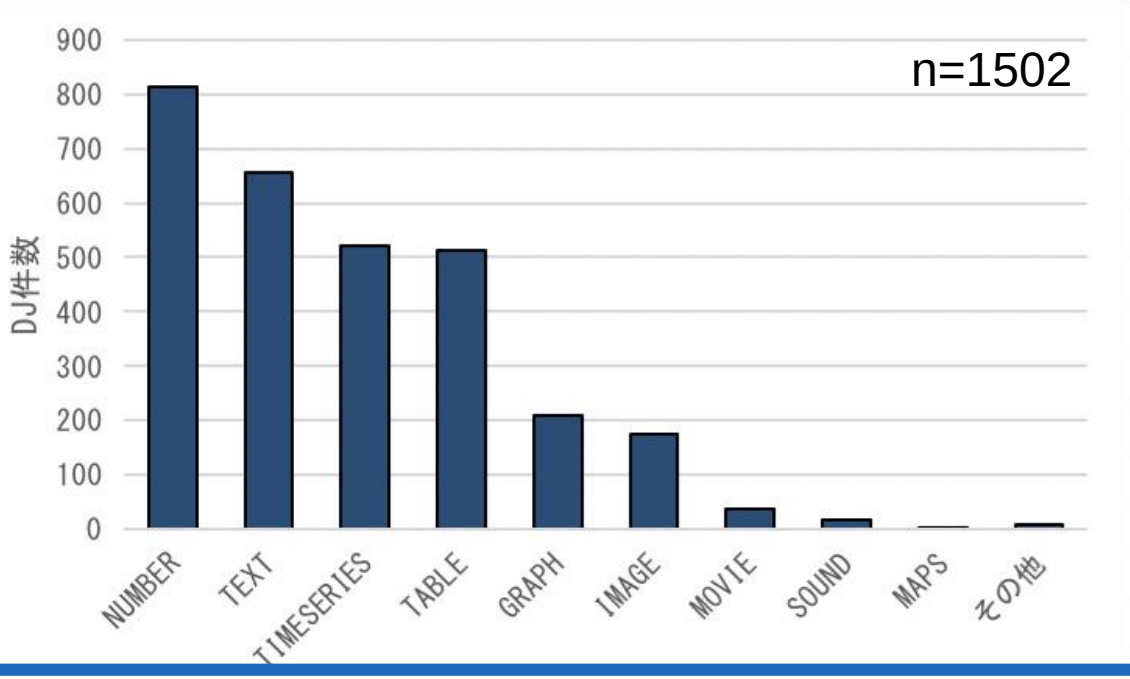
RDF

AED設置場所、Q&Aサイトの投稿とアクセスデータ

JPEG

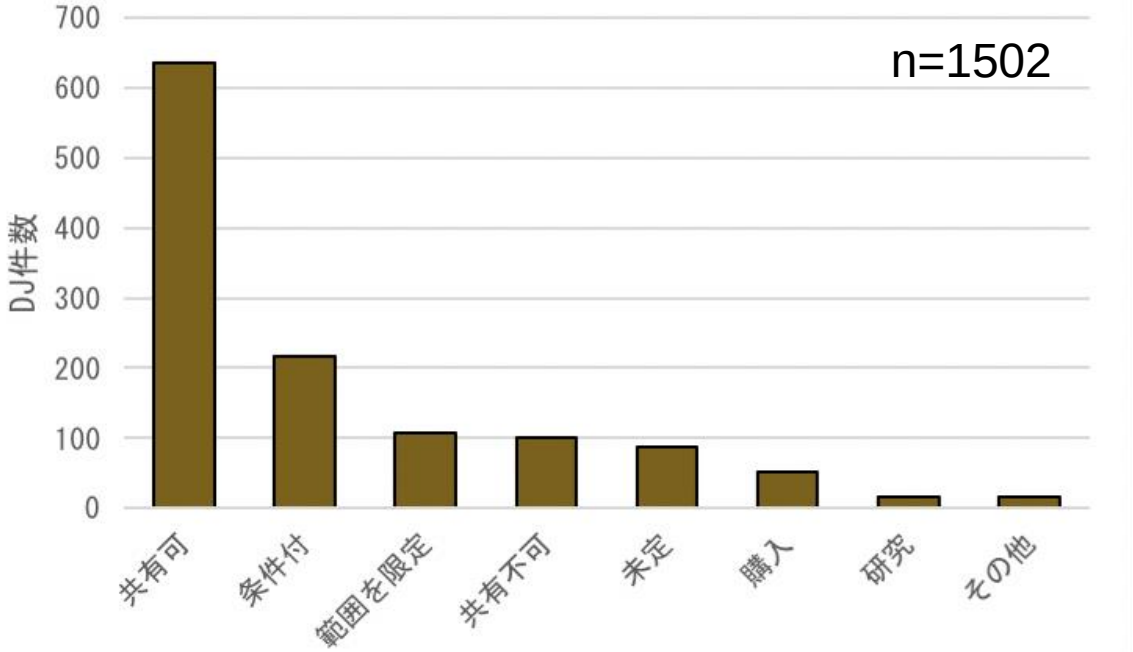
富士山からの日の出写真データ

# DJから分かるデータの属性と特徴




| | |
|---|---|
| NUMBER | GRAPH |
| 飲食店の来店者嗜好データ | つくばセンター放射線測定結果 |
| TEXT | IMAGE |
| 顧客からの問い合わせデータ | 地球上の重力分布図 |
| TIMESERIES | MOVIE |
| 運転時のタイヤの回転ログ | ストリーミングによるIR動画、脳情報のコーディング |
| TABLE | SOUND |
| 国別人口分布の表 | IMDJワークショップの会話ログ |

# DJから分かるデータの属性と特徴




**共有可**

オープンデータ、研究成果（論文）など

**条件付**

手術室内移動データ、dアニメストアのウィークリーランキング情報

**範囲を限定**

原油生産事業データベース、秋田美人に関するデータ

**共有不可**

自動車の点検履歴データ、マス広告のTRP時系列

**未定**

陸上(短距離)選手の強化手法データ、首都圏の鉄道乗車人数

**購入により共有可**

オンライン記事、Twitterのテキスト、消費者の特定商品購入実績データ

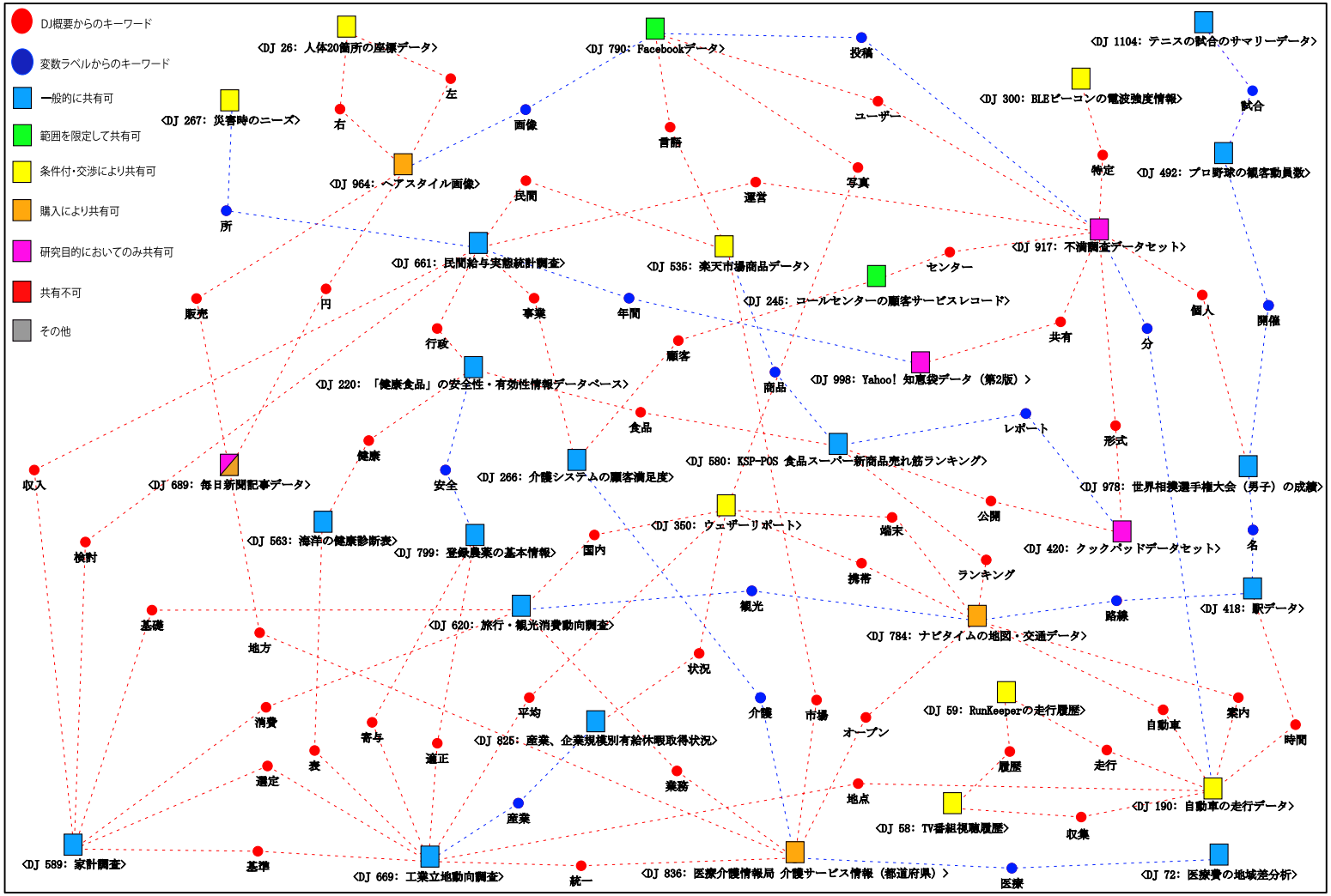
**研究目的**

ニコニコ動画コメントデータ、特許公開情報に基づくテキストデータ

# データ可視化

# DJ間の関係性を可視化

DJの可視化マップを通して、データの潜在的な組み合わせ可能性を議論できる（e.g., KeyGraphを用いて可視化）

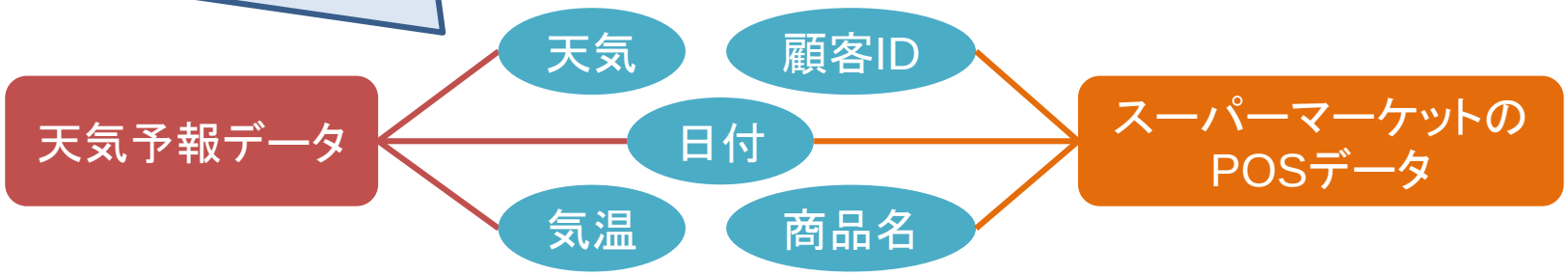


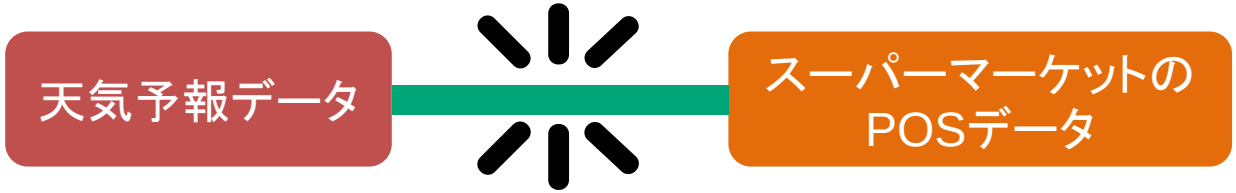
# DJのネットワーク分析

データの結合・連携：**データに含まれる変数を揃えることによって達成**

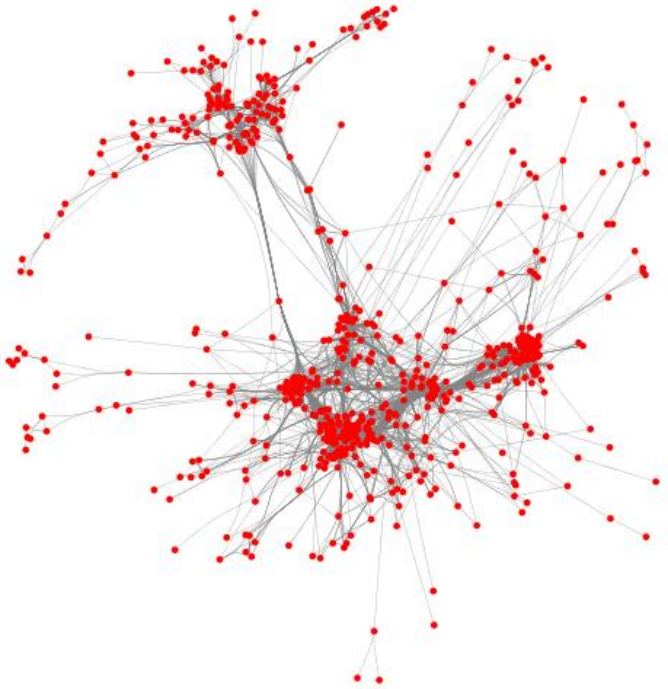
**共通する変数ラベルを有しているデータ同士は結合可能性が高い**

各ノードをDJとし、DJノード間を繋ぐリンクは**DJ同士が共通する変数ラベルを保有**している場合に張られるものとする



# 変数ラベルを介したDJのネットワーク



| ネットワーク特徴量 | 値 |
|---|---|
| リンク数 | 11077 |
| ノード数 | 652 |
| 平均次数 | 33.98 |
| リンク密度 | 0.0522 |
| クラスタ係数 | 0.703 |
| 同類選択性 | 0.561 |
| ネットワーク直径 | 11 |
| 平均経路長 | 3.442 |

**クラスタ係数**が高く、**リンク密度**が低い

- 局所的に距離が近く、大局的には疎なネットワーク
- 同じようなデータ同士は密なネットワークを形成

**平均経路長**が短い、**同類選択性**が高い

- 自然界のネットワークと比較し、人間関係のネットワークに近い
- スモールワールド性を有する

# DJネットワークの中心性

## 次数中心性

あるノードが他のノードとどの程度繋がっているのかを示す指標

| DJタイトル | 次数中心性 |
|---|---|
| Facebookデータ | 0.192 |
| 静岡県三島市の公衆トイレ情報 | 0.187 |
| Twitterデータ | 0.186 |
| 震源リスト | 0.181 |
| オゾン層に関するデータ | 0.180 |

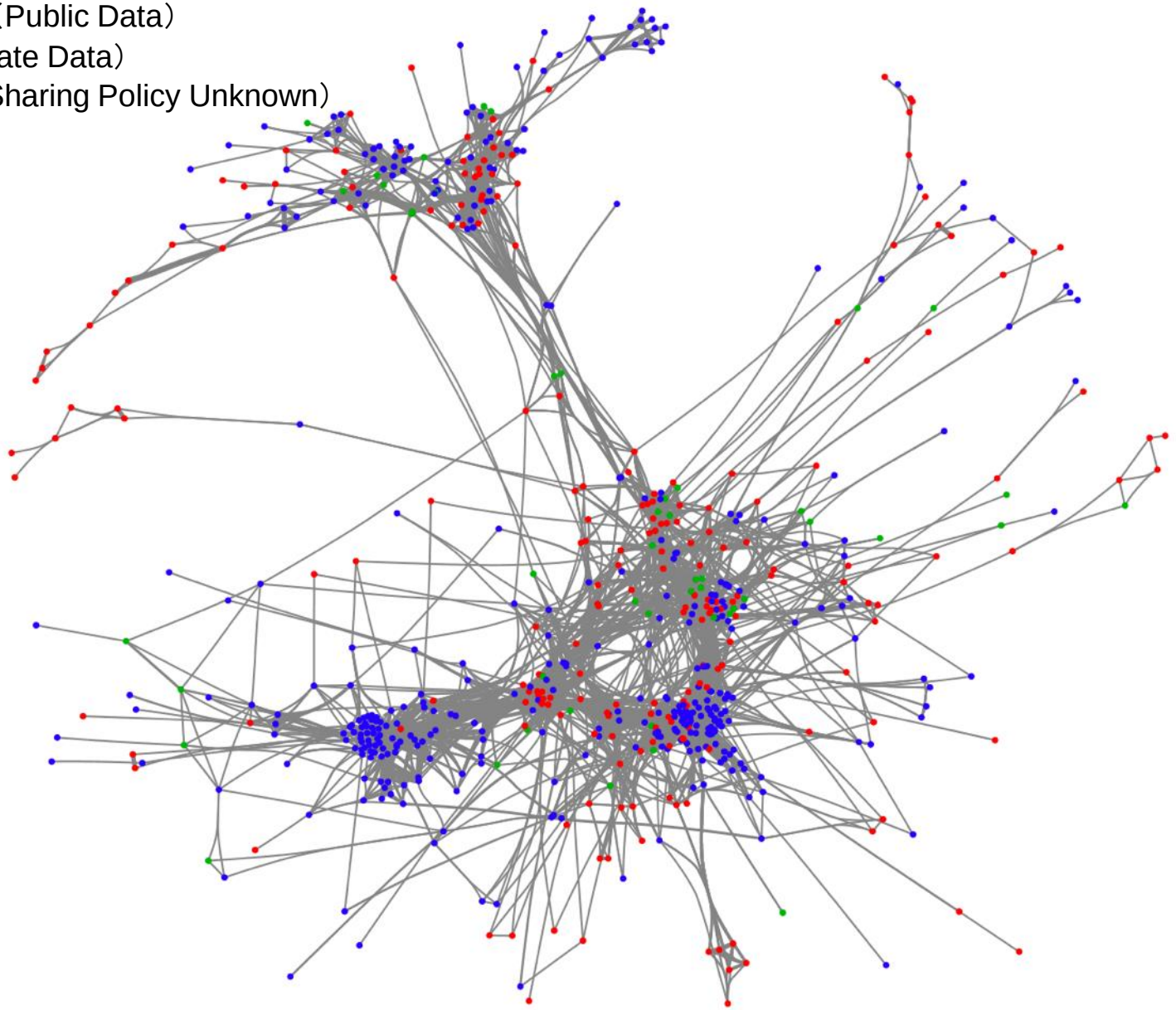
## 媒介中心性

あるノードがどの程度異なる集団を繋ぐかということを測る指標

| DJタイトル | 媒介中心性 |
|---|---|
| 日本の高速道路の交通データ | 0.221 |
| SNSの投稿データ | 0.071 |
| 国・地域別日本食の売上データ | 0.064 |
| Happiness around the World | 0.054 |
| 「読書メーター」データ | 0.035 |

# DJを用いたデータランドスケープ

●:共有可能データ(Public Data)
●:秘匿データ(Private Data)
●:共有条件不明(Sharing Policy Unknown)



# データ検索

# ユーザーはどのようにデータを検索する？


震災後の避難所における被災者の状況を知りたい

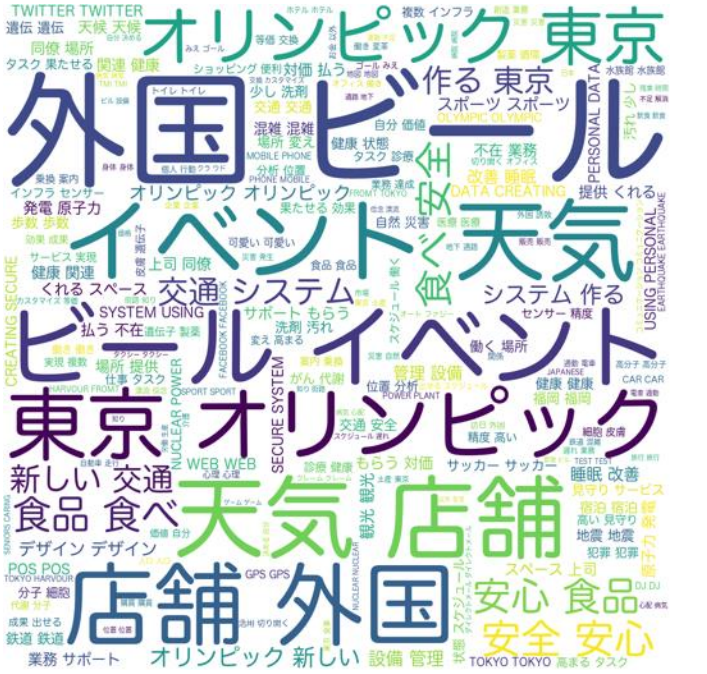
データ検索のクエリは具体的ではない

ユーザーが真に欲しているデータが発見できない

約10,000件のデータ検索クエリを分析

- 会社選択時の優先項目
- 介護施設の設計、効率化と品質改善
- 運動能力 データ
- 安全と生活の国際比較
- リスク 食　安全
- インバウンド外国人 国籍
- 学習意欲 データ
- 楽しい散歩
- 幸せになりたい
- 季節に影響を受けないデザートを作る
- 健やかな子どもの成長

などなど

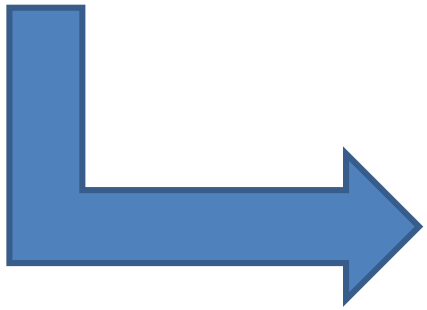


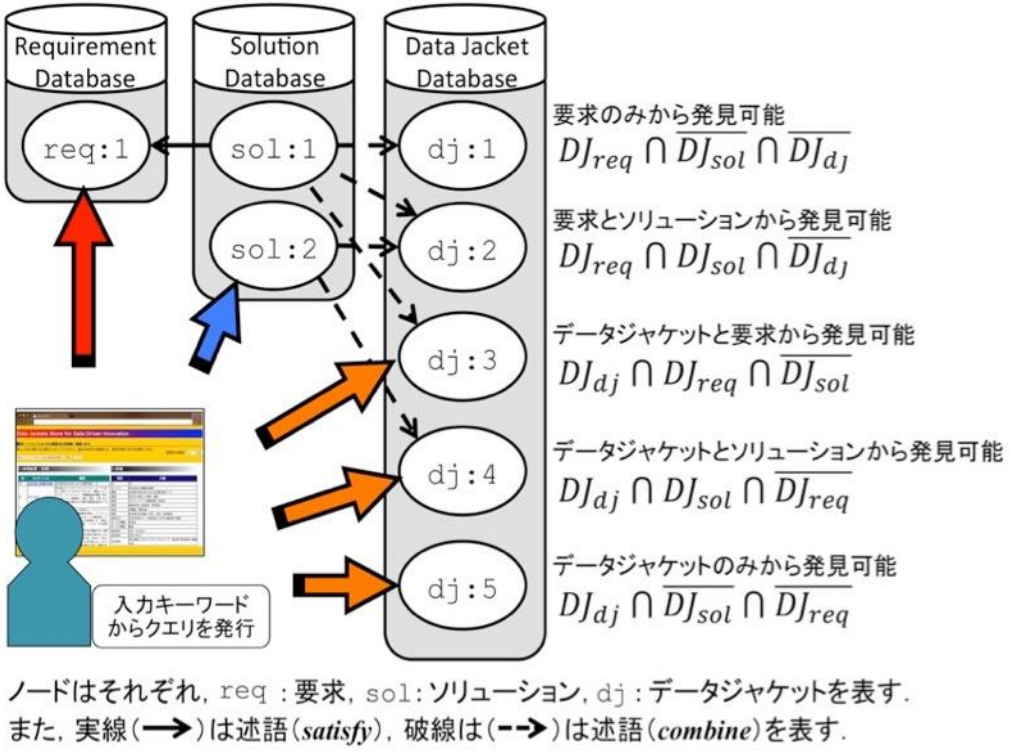
# データ利活用知識の構造化と検索システム

（早矢仕・大澤, 2016）

「要求」とその解決方法「ソリューション」の構造化と再利用

IMDJで創出された「要求」及び「ソリューション」





# データ利活用方法の検討

# データ利活用方法検討ワークショップの手法

## Innovators Marketplace on Data Jackets（IMDJ）

- データ利活用に対する人間の創造性とデータの価値発見を支援するワークショップ手法（Ohsawa et al., 2013）
- 参加者同士のソリューションの評価により、データ保有者は自身のデータを公開することなく活用方法を知り、利用者と取引に関する交渉を開始できる。

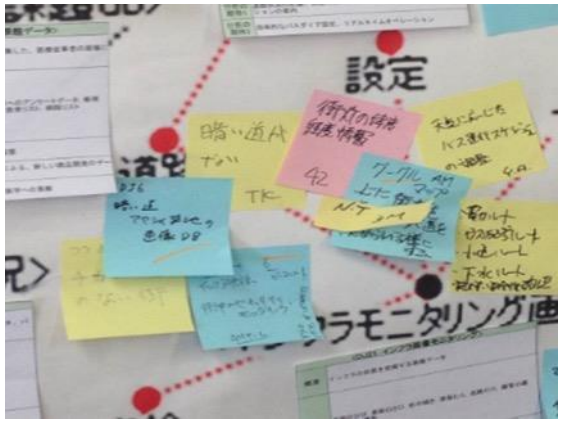
保有者:
DJを提供

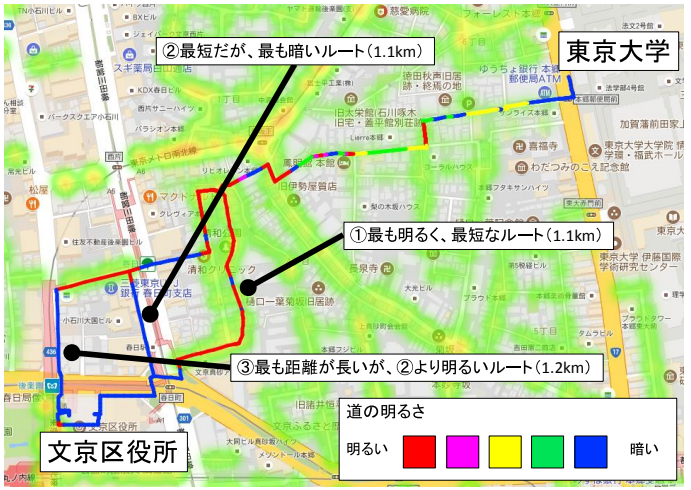
利用者:
要求を提起

提案者:
利活用案を提案

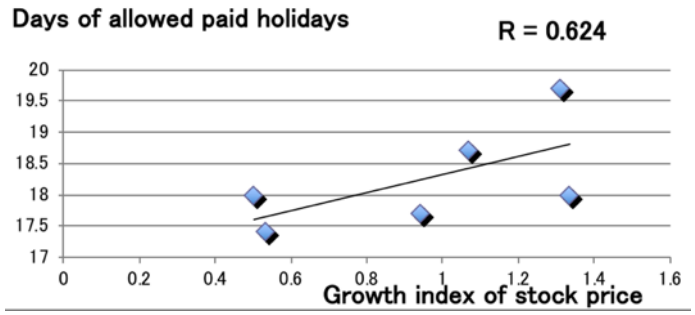
# データ利活用事例の一部（1/2）

## 街路灯データ＋地図データ



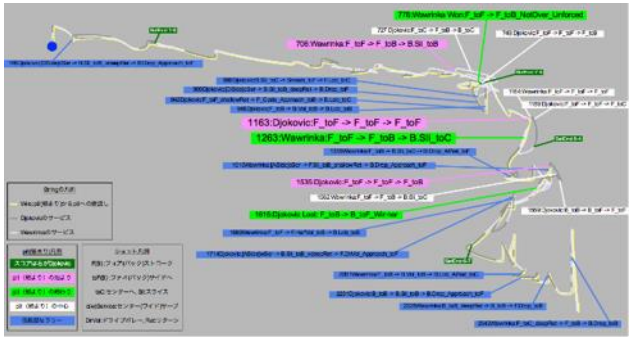




## 有給休暇取得率＋株価データ



# データ利活用事例の一部（2/2）

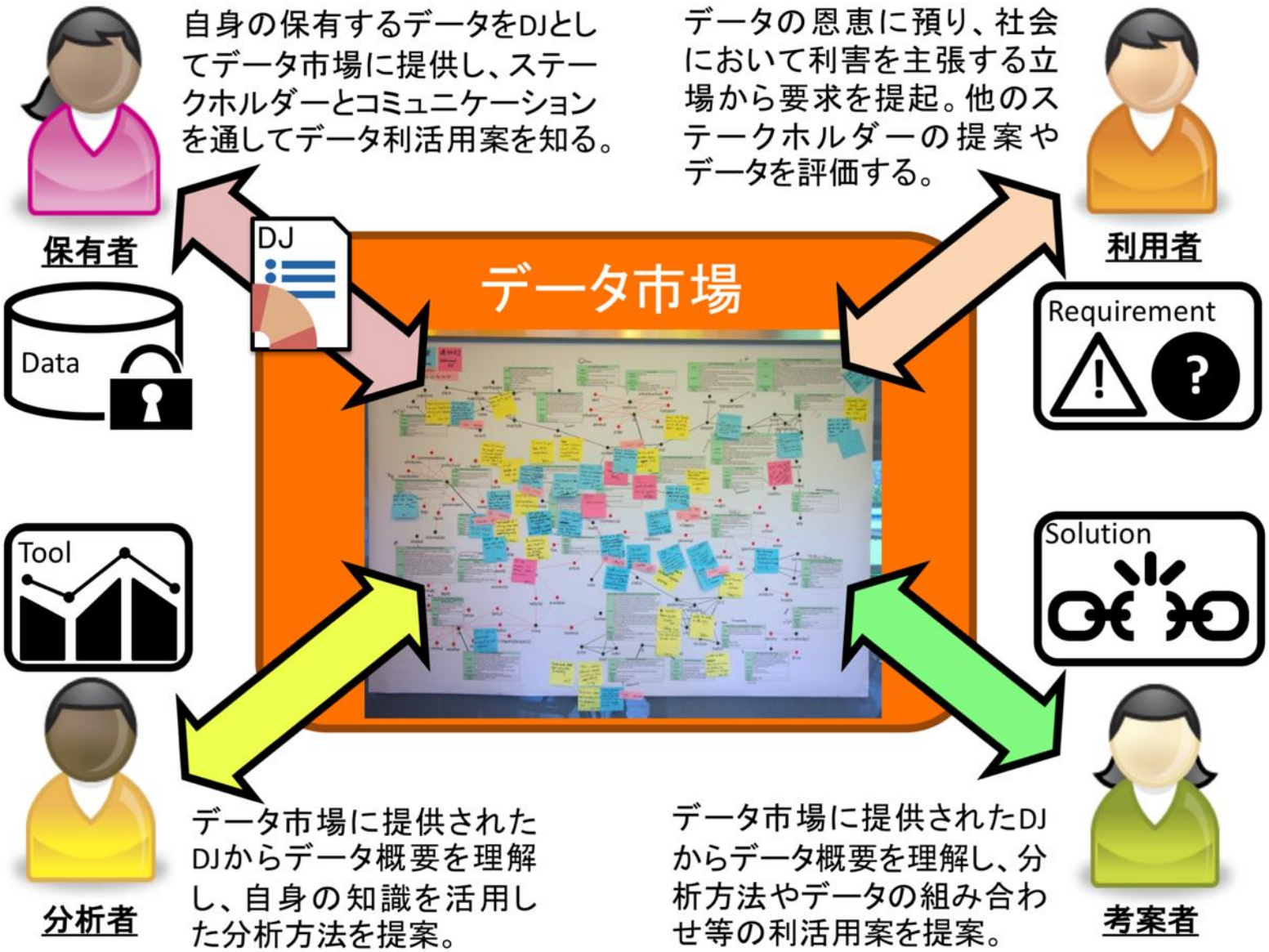
テニスのプレー推移の可視化と分析




自転車の走行履歴から、危険箇所の可視化


サッカーコーチの支援システム

# データ保有者と利用者の間のギャップの解決

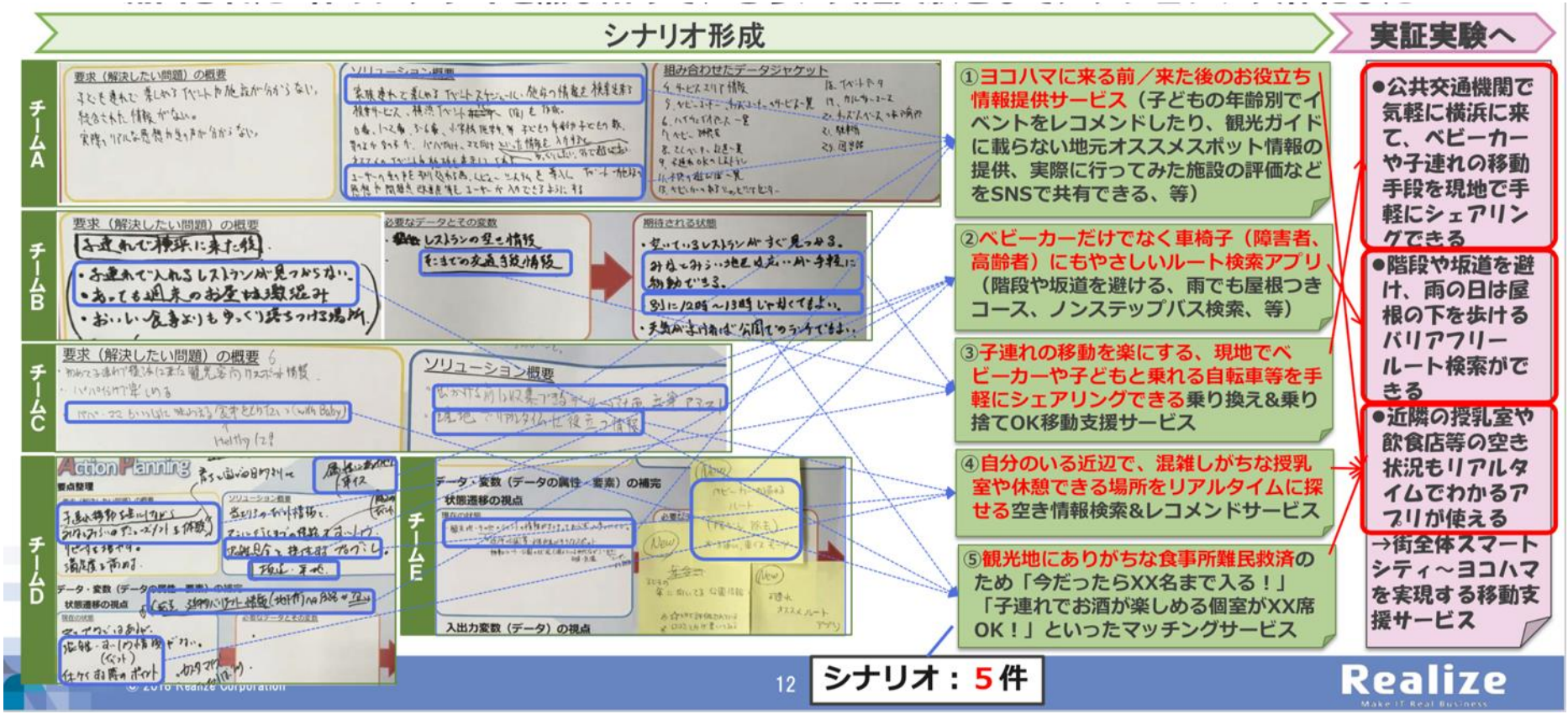


# 横浜市の異分野データ連携プロジェクト

データを活用した課題解決のための公民共創の場を創出

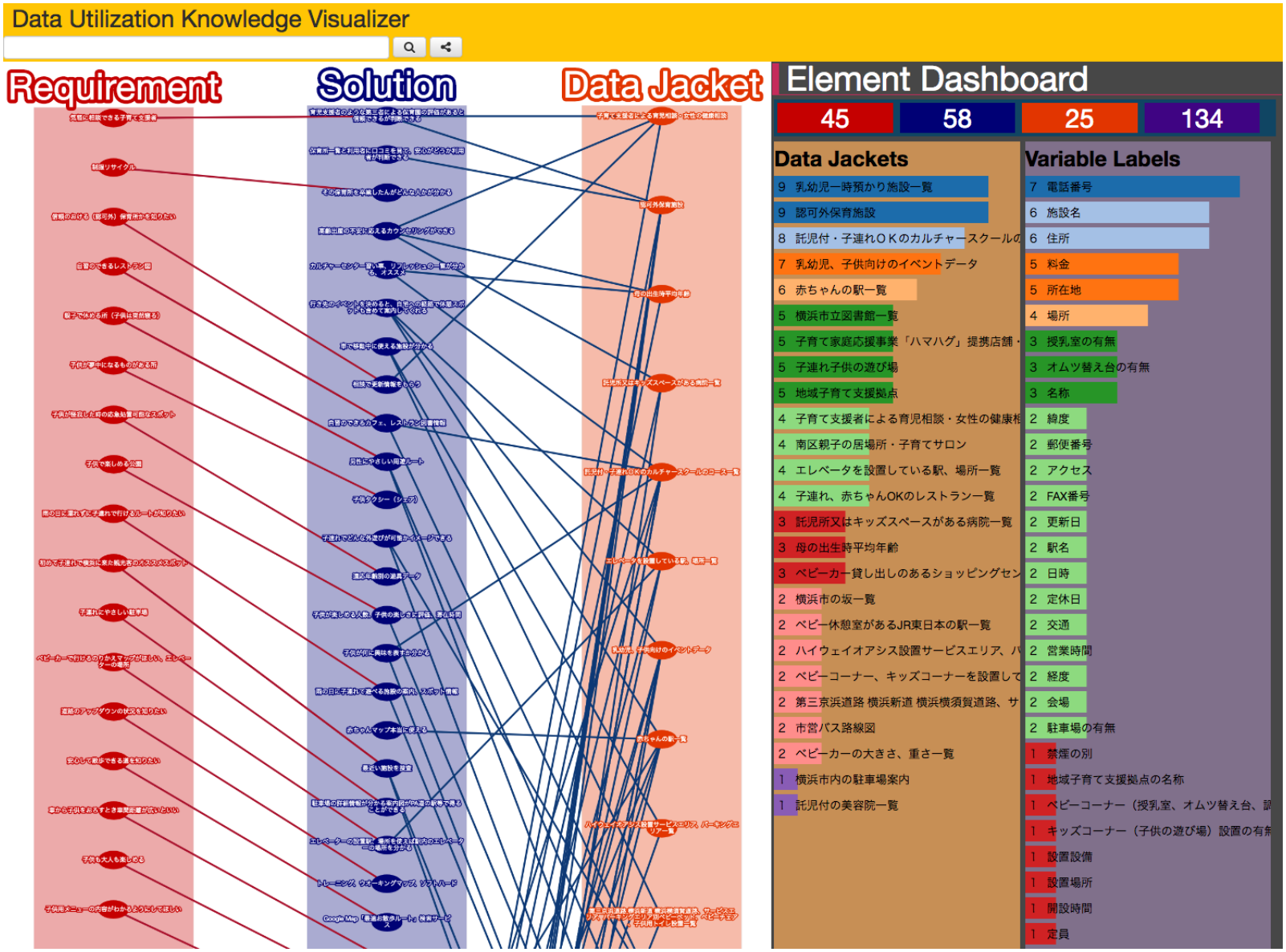

行政情報システム研究所，株式会社リアライズ：
官民データ活用シナリオ創発プラットフォーム事業資料より

# 横浜市の異分野データ連携プロジェクト




行政情報システム研究所，株式会社リアライズ：
官民データ活用シナリオ創発プラットフォーム事業資料より

# データ利活用知識の構造化と再利用



# まとめ

- データだけには価値はなく、データに文脈を乗せることで初めて「材」から「財」となる
- データに関わる多様な人々の様々な知識・経験をデータに加えることで異分野データ連携が実現

The 1st International Workshop on

# Cross-disciplinary Data Exchange and Collaboration

in ICDM2018


**論文締切：**

**8/7（火）**

**Webページ：**

**http://www.panda.sys.t.u-tokyo.ac.jp/CDEC/2018/**