ビッグデータの活用事例と分析方法

佐藤彰洋
京都大学大学院情報学研究科
科学技術振興機構さきがけ
aki@i.kyoto-u.ac.jp

目的

- なぜデータを分析する必要があるのか？

- 何かを決めたいから（意思決定）

Big Dataとは？

- ビッグデータ： McKinsey Global Institute (MGI)による”Big data: The next frontier for innovation, competition, and productivity”

3V

Volume（量）, Velocity（速度）, Variety（多様性）

Big Data の 3V + 4V

- **Volume (量)**
- **Velocity (速度)**
- **Variety (多様性)**

- **Variability(可変性)**: データの構造、内容、質などの特性が変化する程度
- **Value (価値)**: 人的・金銭的コストに対して得られる便益の大きさ、または、その期待値
- **Veracity (信頼性)**: データ品質と関係
- **Volatility (変動性)**: データが価値を有する時間（価値と速度の複合概念）

OECD (経済協力開発機構)

“Data-driven innovation for growth and well-being”

- イノベーションと成長および福祉を促進するためのデータの役割に関する分析と, 知識に基づく資産(Knowledge-Based Capital: KBC)に関する分野横断的なプロジェクトの位置づけ
- 成長と福祉を促進するためのデータの役割に関するエビデンスの改善及び、データ駆動型経済(data-driven economy)の便益の最大化とこれに関連するリスクの低減を行うために必要とされる政策ガイドライン

OECD: Data-driven innovation for growth and well-being

- 科学的発見のためのデータの役割
- 健康向上のためのデータの役割
- より良い統治を行うためのデータの利活用
- クラウドコンピュータ、分析、および他の技術的可能性
- 技能と雇用に関するほかの関連性
- データ駆動型社会のための信頼性確保
- 知識ベース資産としてのデータへの投資尺度

モットー

我々は測れないものは管理できない

政府統計

- e-Stat 政府統計の総合窓口

https://www.e-stat.go.jp

政府統計の所管省庁

- **内閣官房**
- **人事院**
- **内閣府**　公正取引委員会、警察庁、消費者庁
- **総務省**　公害等調整委員会、消防庁
- **法務省**
- **外務省**
- **財務省**　国税庁
- **文部科学省**　文化庁
- **厚生労働省**　中央労働委員会
- **農林水産省**　林野庁、水産庁
- **経済産業省**　資源エネルギー庁、特許庁、中小企業庁
- **国土交通省**　観光庁、海上保安庁
- **環境省**
- **防衛省**

基幹統計調査

医療施設調査，科学技術研究調査，家計調査，学校基本調査，学校保健統計調査，学校教員統計調査，患者調査，海面漁業生産統計調査，ガス事業生産動態統計調査，漁業センサス，牛乳乳製品統計調査，経済産業省企業活動基本調査，経済産業省生産動態統計調査，経済産業省特定業種石油等消費統計，経済センサス－基礎調査，経済センサス－活動調査，建築着工統計調査，建設工事統計調査，小売物価統計調査，国民経済計算，国民生活基礎調査，国勢調査，個人企業経済調査，工業統計調査，鉱工業生産・出荷・在庫指数，港湾調査，作物統計調査，産業連関表，住宅・土地統計調査，就業構造基本調査，社会生活基本調査，社会教育調査，人口動態調査，社会保障費用統計，商業統計調査，商業動態統計調査，自動車輸送統計調査，生命表，全国消費実態調査，全国物価統計調査，生産動態統計，生産能力指数・稼働率指数，石油製品需給動態統計調査，船員労働統計調査，造船造機統計調査，地方公務員給与実態調査，賃金構造基本統計調査，鉄道車両等生産動態統計調査，特定サービス産業実態調査，内航船舶輸送統計調査，農業経営統計調査，農林業センサス，法人企業統計調査，法人土地・建物基本調査，毎月勤労統計調査，埋蔵鉱量統計調査，民間給与実態統計調査，木材統計調査，薬事工業生産動態統計調査，労働力調査

人間の能力

- **人間の脳の記憶容量はどの程度か？**　　数PB～100GB
- **人間の応答速度はどの程度か？**

　　人間の瞬間は50ms（これ以下の変化は連続とみなす）

- **人間の記憶速度はどの程度か？**　2bps （音声、文字、視覚ともに）
- **人間の一生はどの程度か？**

80年間＝80years x365 days x1440 minutes x60 seconds=2.5x10^9s=25億秒

- **人間が一生に覚えることのできる量の上限はどの程度か？**

2bps x 2.5 x 10^9s × 0.5 (利用可能時間)= 2.5 Gbits = 312.5MB

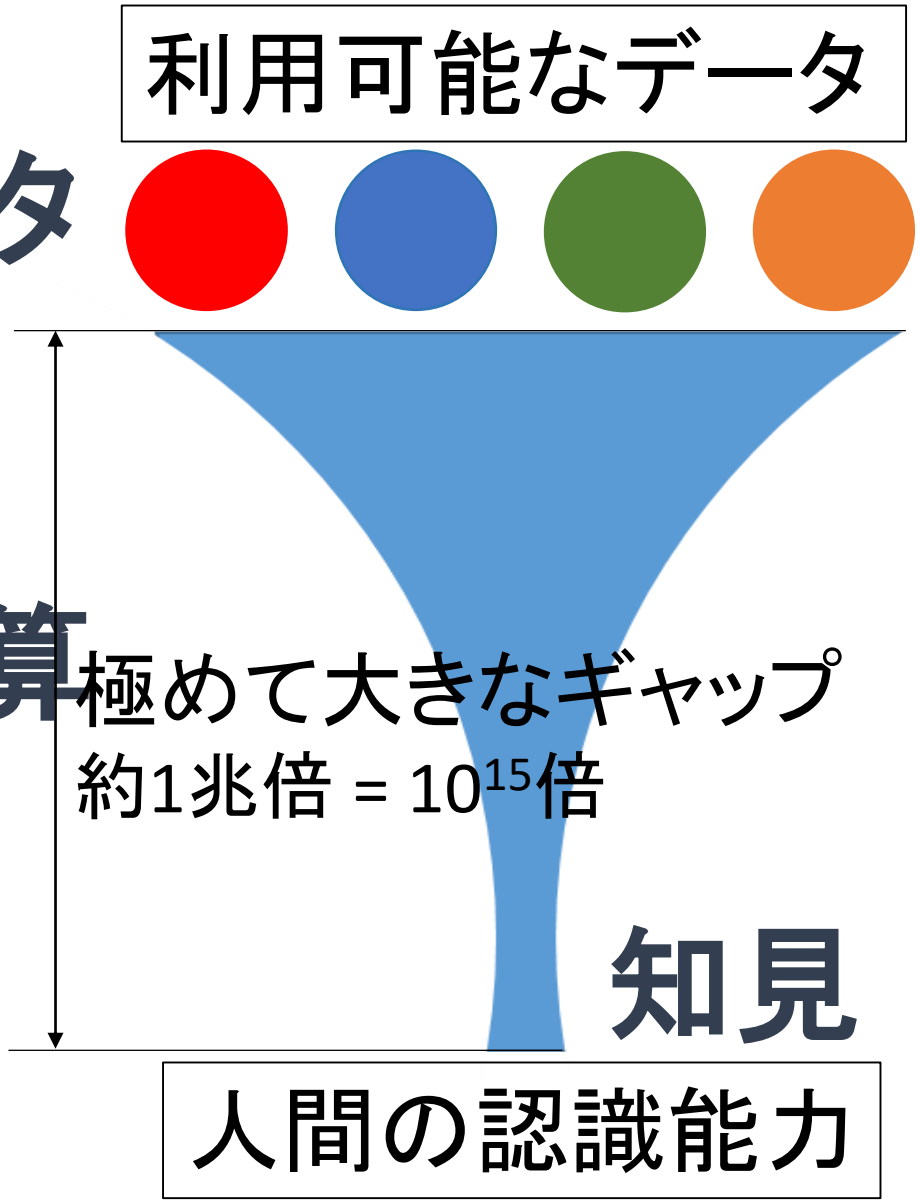
- **人間が一生にすることができる物事の量はどの程度か？**

データ基盤

約 44ZB/year by 2020　(ZB = 10^{21} Bytes)

1. 専門知識の特定
2. 目的の特定と共有
3. データ源の特定
4. 計算方法の特定
5. 計算資源の特定
6. 計算手順の特定
7. データ分析の実行
8. 計算結果のレポーティング
9. 施策実施方法の特定
10. 施策の実施
11. 実施結果の評価

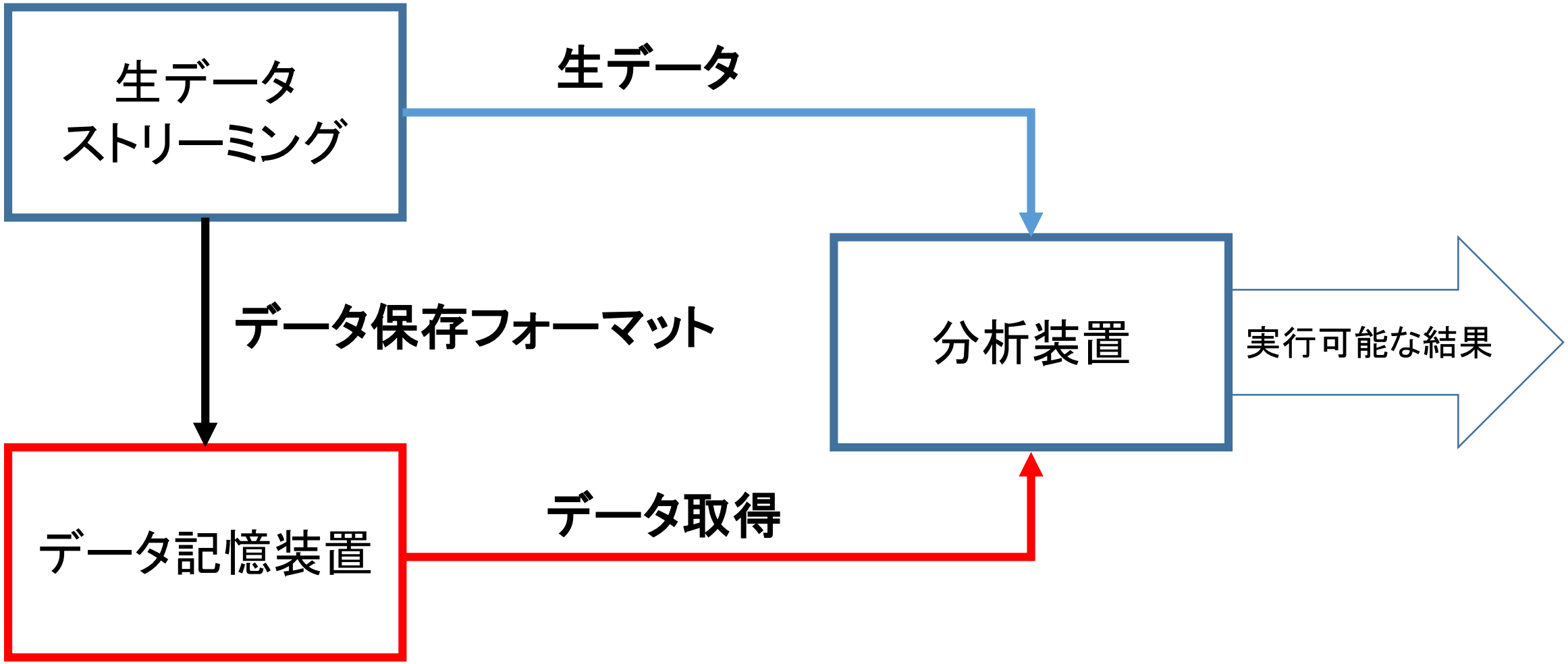
The Digital Universe of Opportunities, Accessed on 18 June, 2017 [ONLINE] https://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm



T.K. Landauer, Cognitive Sci. 10, 477–493 (1986), DOI:10.1207/s15516709cog1004_4

5~10MB/year (MB = 10^6 Bytes)

データ分析の概念図



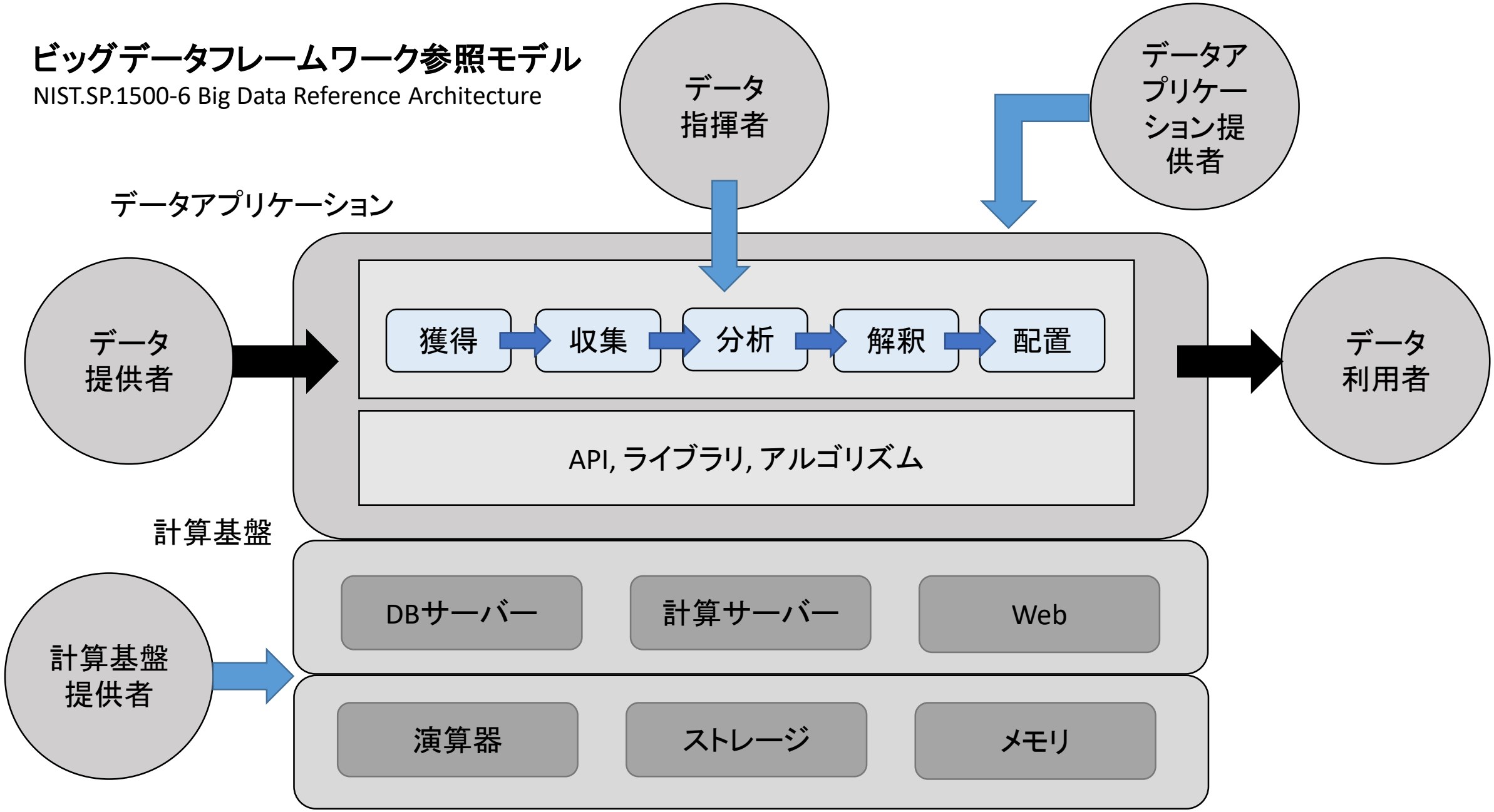
ISO/IEC JTC 1/SC 32 Data Management and Interchange

参照モデルにおける役割

- データ消費者 (data consumer)
- データ提供者 (data provider)
- データ指揮者 (data orchestrator)
- データアプリケーションプロバイダー (data application provider)
- データフレームワークプロバイダー (data framework provider)

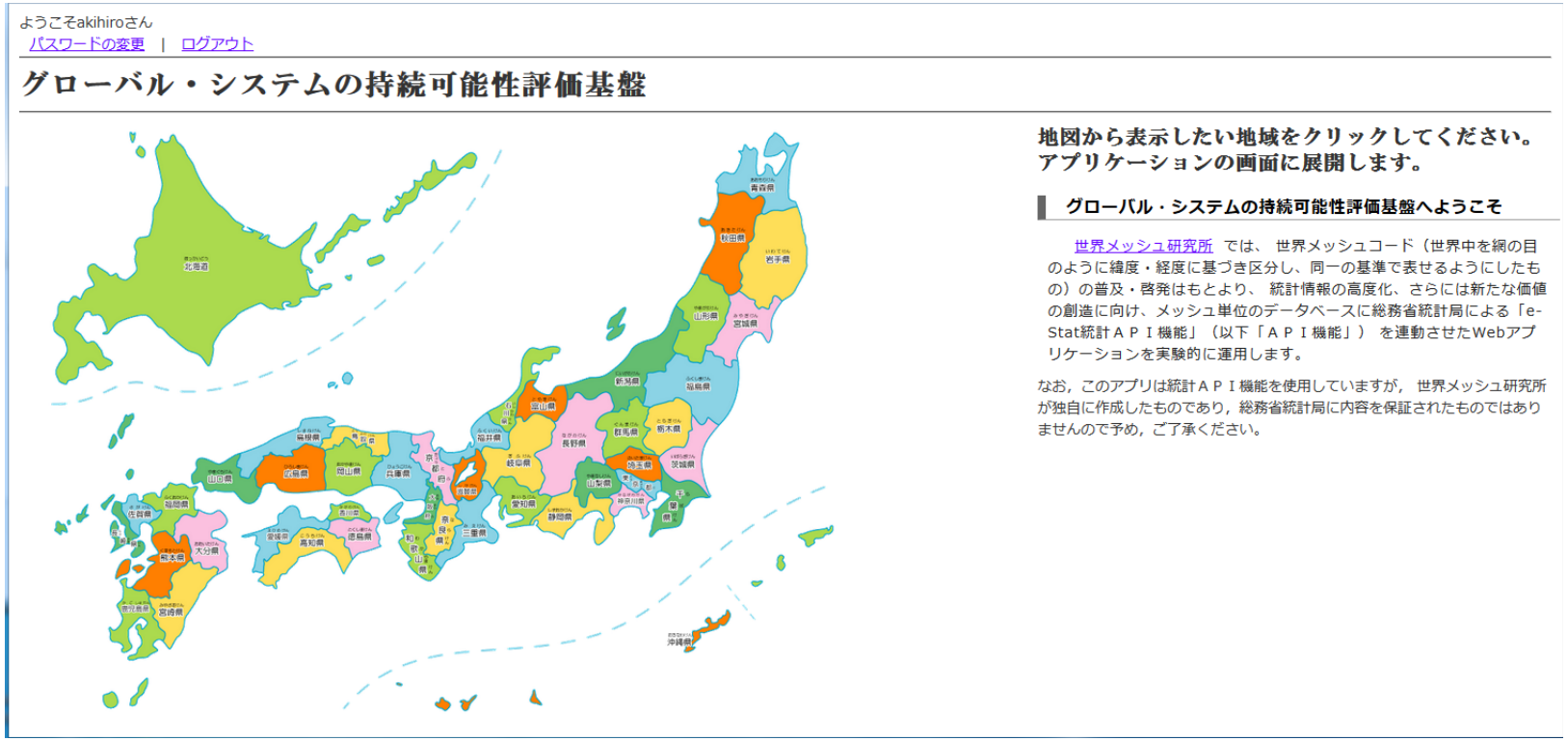
ビッグデータフレームワーク参照モデル

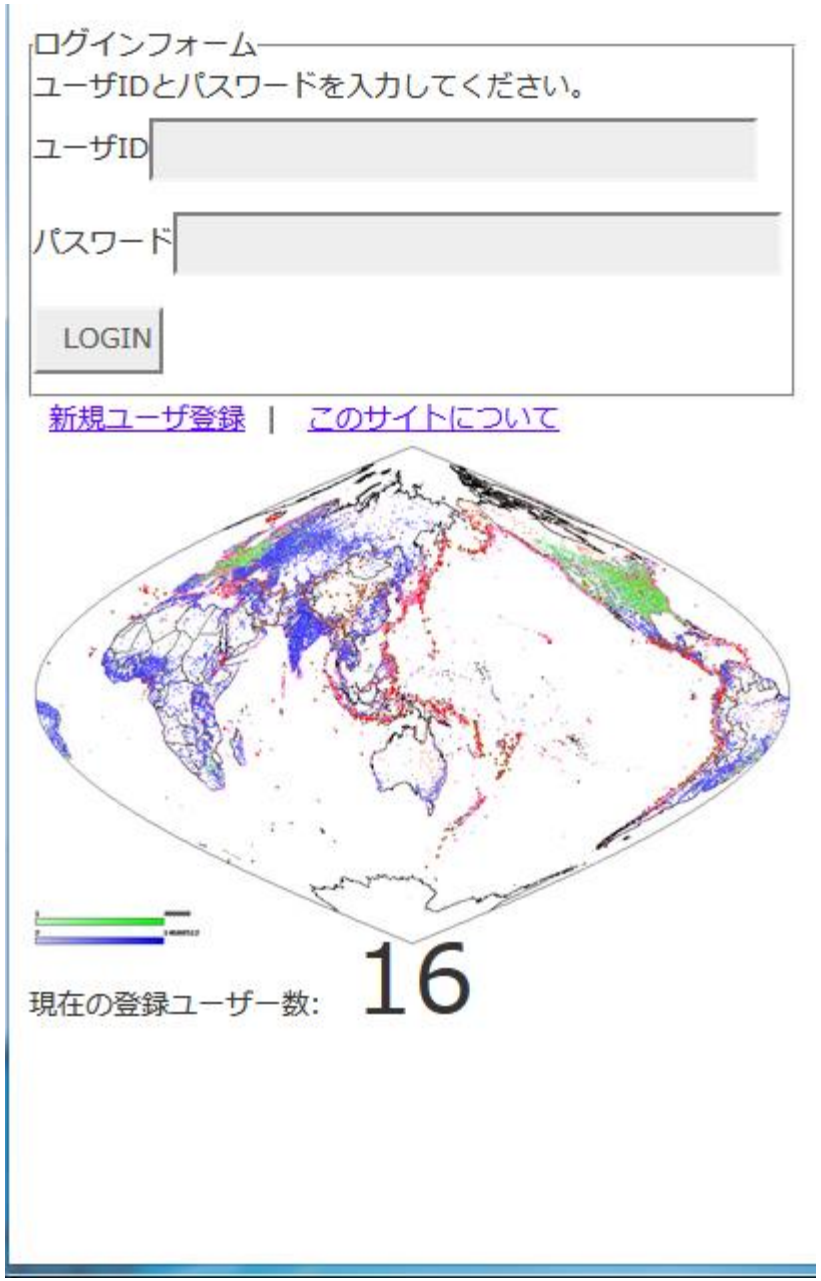
NIST.SP.1500-6 Big Data Reference Architecture



Web Application

http://www.meshstats.xyz/meshstats/login.php





現在提供中の機能

- **利用可能な言語**：英語、日本語、イタリア語、ドイツ語、スペイン語、韓国語、ベトナム語、中国語、ポーランド語

9言語

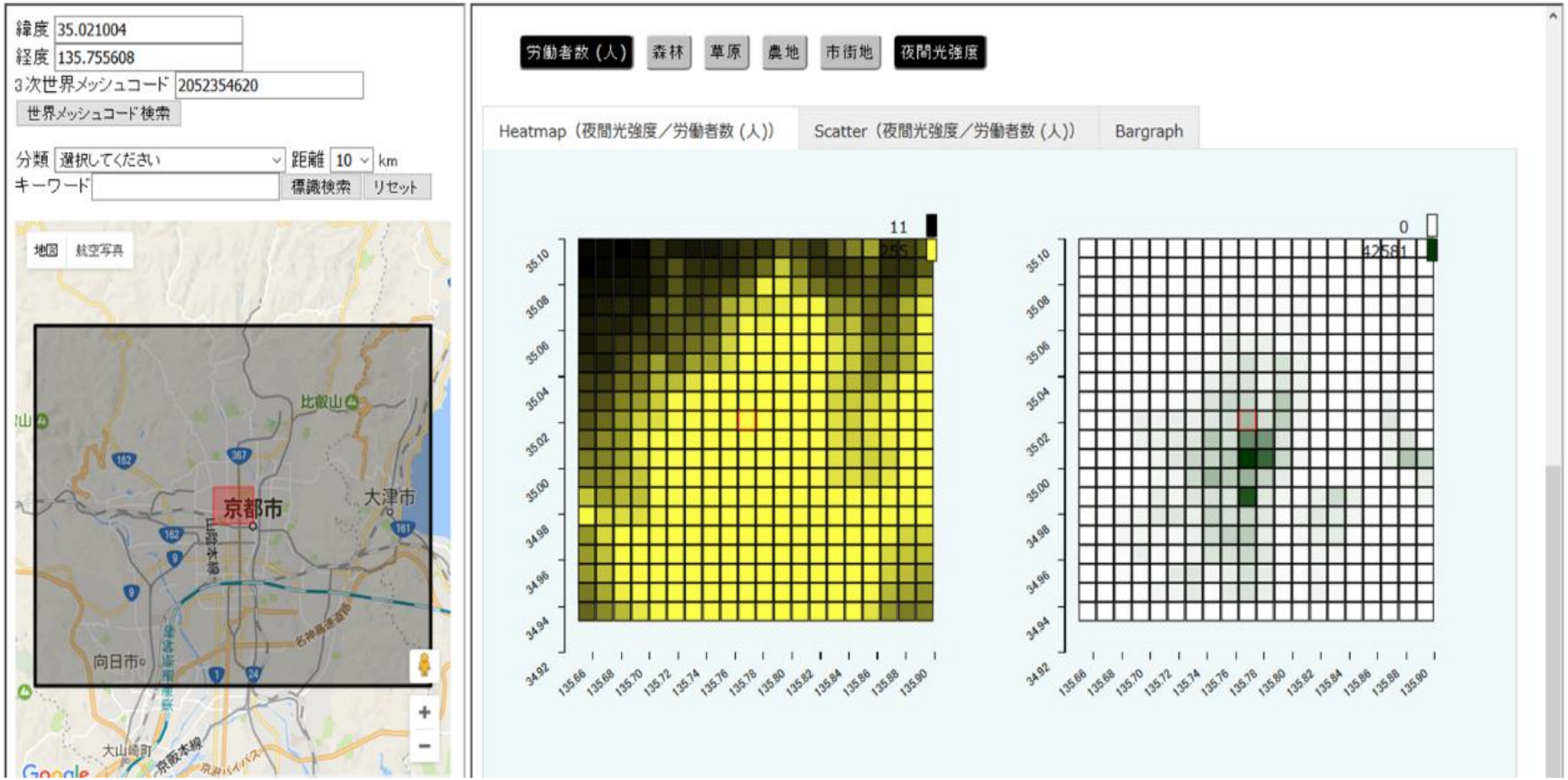
- **利用可能な機能**：基本メッシュ、津波ハザードメッシュ、地震ハザードメッシュ、浸水想定区域浸水深メッシュ、バイオマス森林成長量メッシュ、求人メッシュ、求人時系列、求人賃金メッシュ、求人賃金時系列、近接ホテル探索、宿泊予想メッシュ、月次宿泊実績メッシュ、日次宿泊実績メッシュ、日次宿泊平均価格メッシュ、日次宿泊売上メッシュ、国土交通省宿泊旅行統計メッシュ、国土交通省宿泊旅行邦人統計メッシュ、国土交通省宿泊旅行外国人統計メッシュ、地域観光資源情報 市区町村別月次宿泊実績統計、市区町村別集計鉄道乗降客数統計、鉄道駅周辺宿泊旅行統計、鉄道駅周辺求人状況 鉄道駅周辺宿泊状況、近隣港検索、国際線離発着、国内線離発着

25機能

画面の例


グローバル・システムの持続可能性評価基盤
緯度 35.021004
経度 135.755608
3次世界メッシュコード 2052354620
世界メッシュコード検索
分類 市役所
距離 10 km
キーワード
標識検索
リセット
地図
航空写真
京都市役所 (Kyoto City Hall)
[ブックマークする]
京都市役所 (Kyoto City Hall)
(緯度, 経度)=(35.011641, 135.768190)
日本海
日本
京都
大津
草津
高槻
京都市
Google
imagery ©2017 TerraMetrics
基本メッシュ
津波ハザードメッシュ
地震ハザードメッシュ
バイオマス森林成長量メッシュ
求人メッシュ
近接ホテル探索
宿泊予想メッシュ
月次宿泊実績メッシュ
日次宿泊実績メッシュ
日次宿泊平均価格メッシュ
日次宿泊売上メッシュ
国土交通省宿泊旅行統計メッシュ
国土交通省宿泊旅行邦人統計メッシュ
国土交通省宿泊旅行外国人統計メッシュ
地域観光資源情報
市区町村別月次宿泊実績統計
市区町村別集計
鉄道駅周辺宿泊旅行統計
鉄道駅周辺求人状況
国際線離発着
国内線離発着
3次世界メッシュコード 2052354611 地震ハザードメッシュ (周辺 10 km)
テーブルの開閉

画面の例


緯度 35.021004
経度 135.755608
3次世界メッシュコード 2052354620
世界メッシュコード検索
分類 選択してください
距離 10 km
キーワード
標識検索
リセット
地図
航空写真
京都市
大津市
比叡山
向日市
労働者数（人）
森林
草原
農地
市街地
夜間光強度
Heatmap（夜間光強度／労働者数（人））
Scatter（夜間光強度／労働者数（人））
Bargraph
11
0

画面（機能）

2052354620

(緯度, 経度)=(35.021004, 135.755608)

津波ハザードメッシュ　地震ハザードメッシュ　浸水想定区域浸水深メッシュ　求人メッシュ

ンユ　月次宿泊実績メッシュ　日次宿泊実績メッシュ　日次宿泊平均価格メッシュ　日次宿泊売上メ

統計メッシュ　国土交通省宿泊旅行邦人統計メッシュ　国土交通省宿泊旅行外国人統計メッシュ　鉄道乗

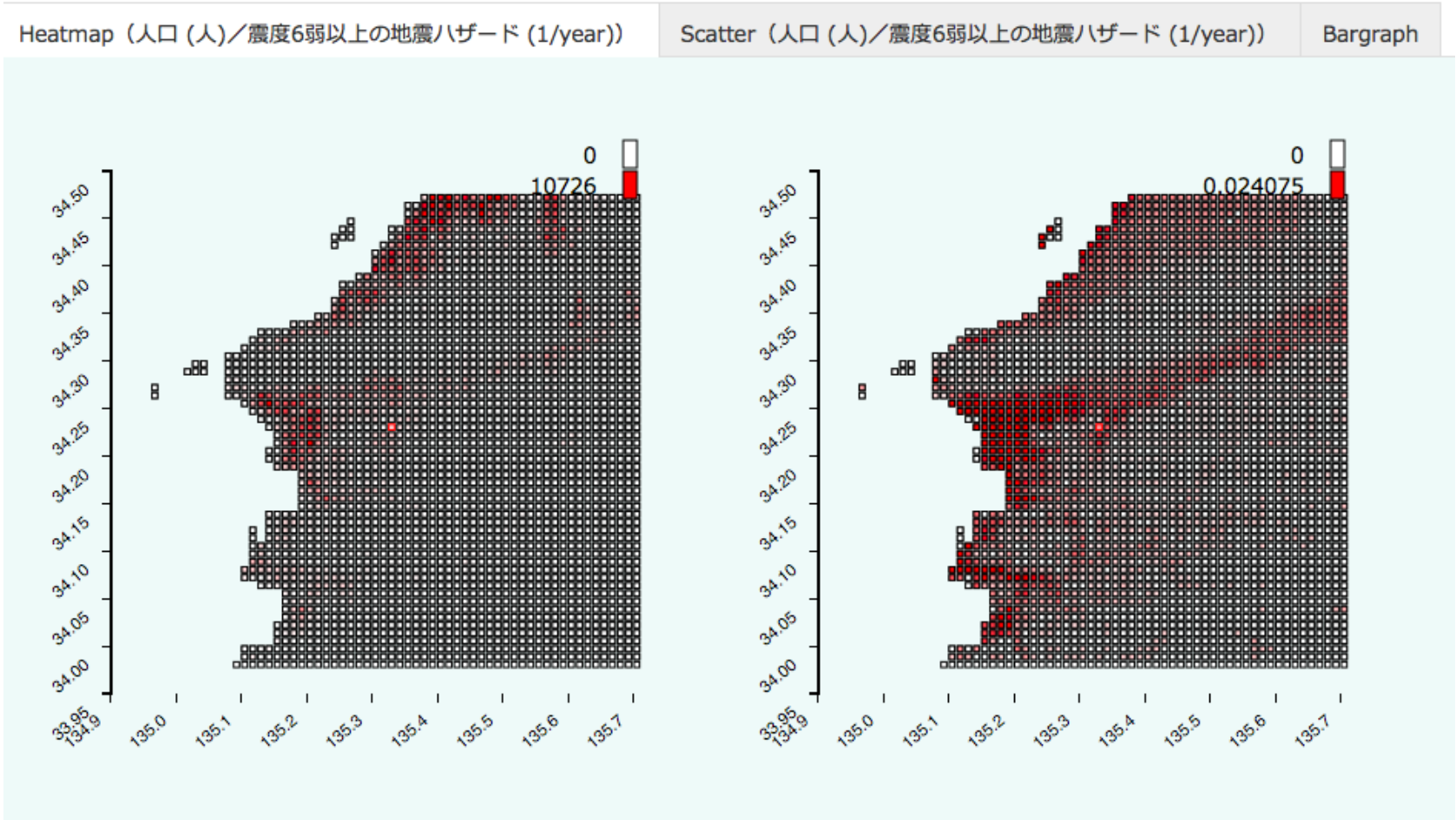
鉄道駅周辺宿泊旅行統計

今回は地震ハザードメッシュを
利用する

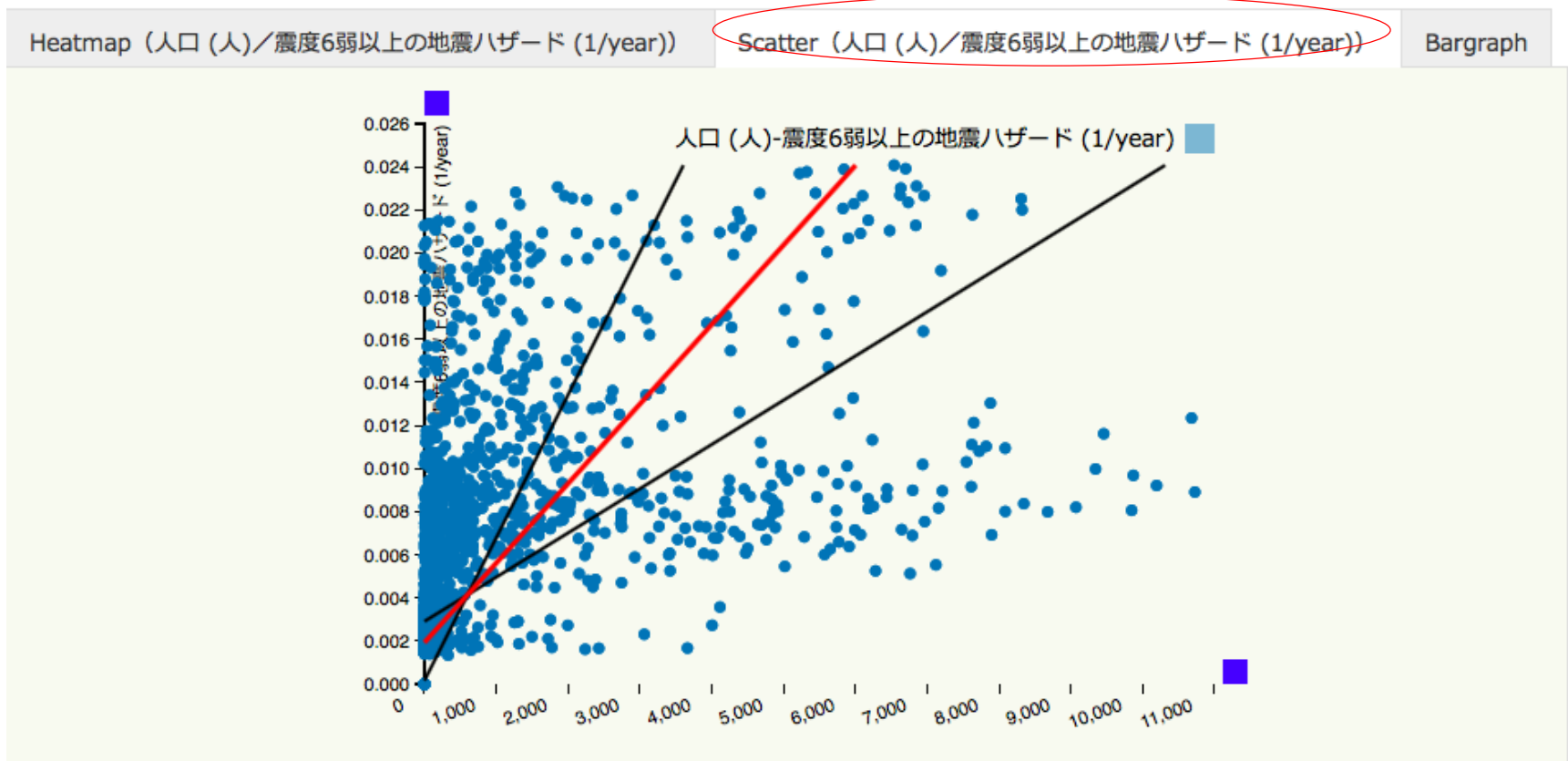
画面（ボタン）

例えば、「人口」と「震度6弱以上のハザード」の関係を調べたい場合、
このように二つ選択する。

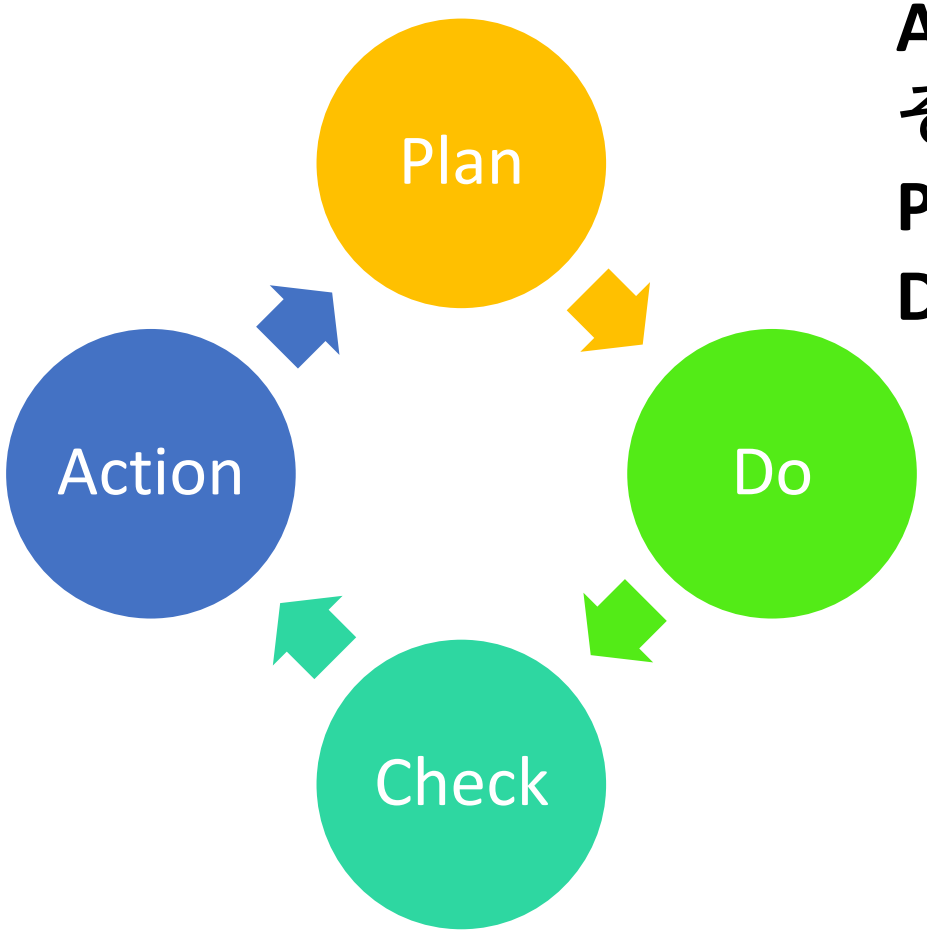
左：人口　右：地震ハザード
このようにそれぞれのメッシュに対して、大きいものほど濃くなるように
表示されている。



ここを選択すると、横軸:人口 縦軸:地震ハザードのグラフが描かれる。

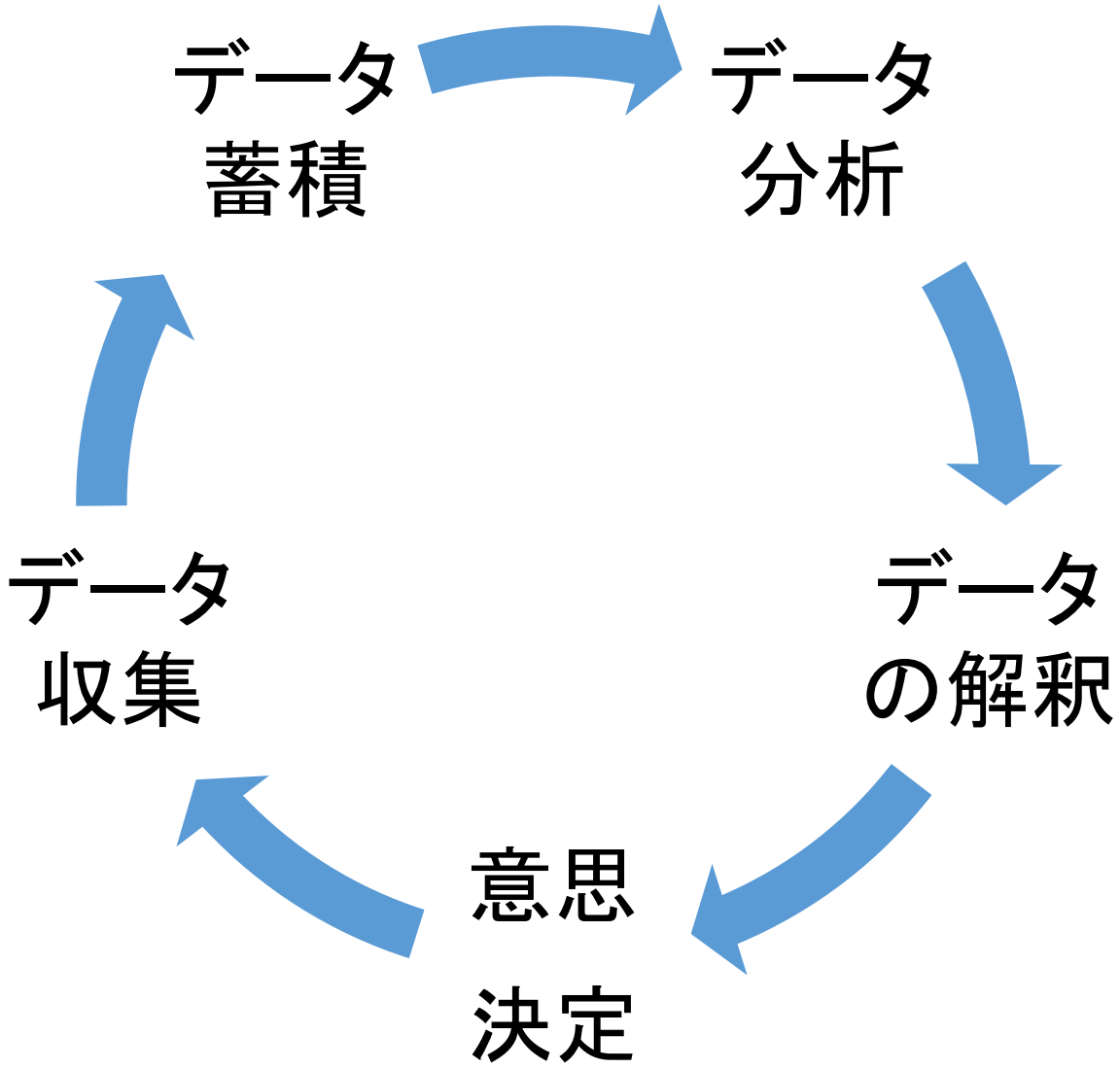


CAPDサイクル

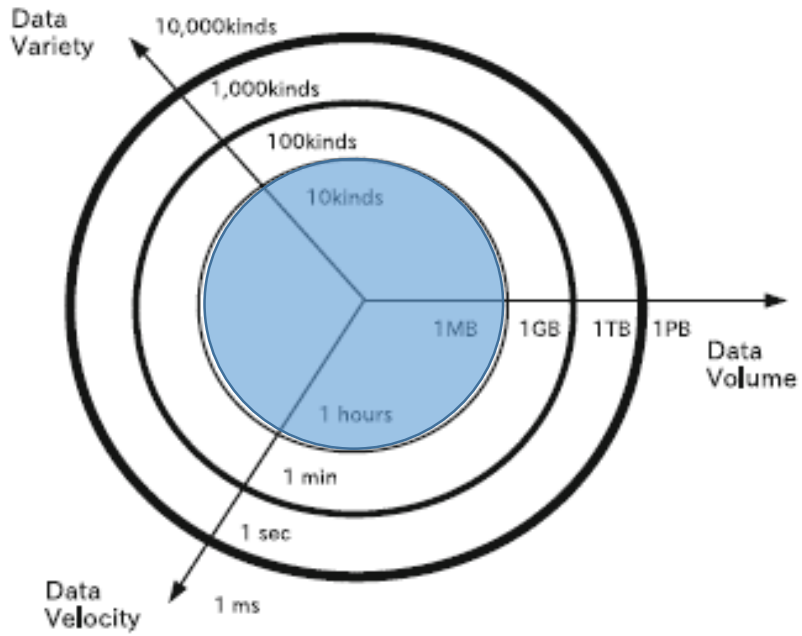


Check: 現在の状況を測定
Action: 望むべき形と現状のギャップを認識し、それが小さくなる意思決定を行う
Plan: 行うべきことの計画を作成
Do: 計画を実施

データ利活用のPDCAサイクル



はじめは小さなデータ断片から初め、スパイラル的に理解を深めつつデータ規模を拡大させていく



参加型デザインサイクル

ブレイン・ストーミング

インプット

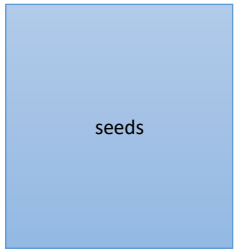
レビュー

プロトタイピング

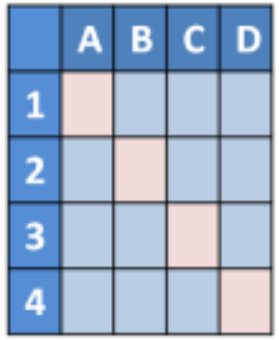
UXデザイン

データ分析

アイデア創出



アイデア拡散と評価



ペルソナ・デザイン

演習

1. 場所の特定：日本国内で我々のグループワークを行う場所と課題となる行政サービスを決めます
2. 組織化：６名でグループを作ります
3. 資源とリスクの特定：決めた場所における“行政サービス”の様子をスケッチしましょう。
4. どのような行政サービスに活かせる資源がありますか？また、その場所で行政サービスを考える場合、どのようなリスクが想定できますか？
5. データ分析による定性的理解から定量的理解
6. 共有