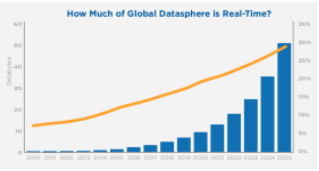
mdx: 大学・研究機関で共創する産学官連携のためのデータプラットフォーム

東京大学 情報基盤センター

田浦健次朗

データ科学・利活用を取り巻く状況

- データが重要な資産（研究、ビジネス、公共政策、など様々なセクターで）



データの増大

オープンサイエンス

DATA.GOV

DATA GO.JP

VLED

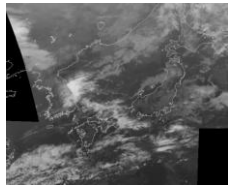
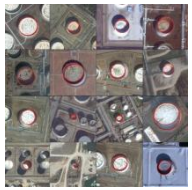
データ活用の重要性への認識の高まり

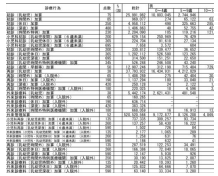
データ

データ科学・活用の潮流・ドライバ

- **機械学習（ML）・AI、ツールの発展（MLの「民主化」）**


- **社会応用に直結するデータの整備**


GPS

気象

衛星

公共交通

医療行為

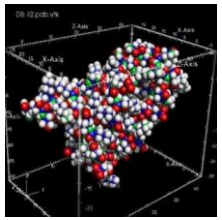
道路・インフラ

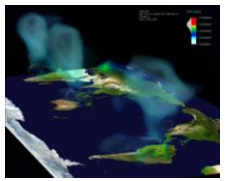
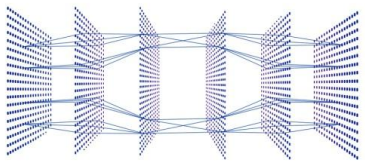
- **センサ（IoT）データ、実時間応用**





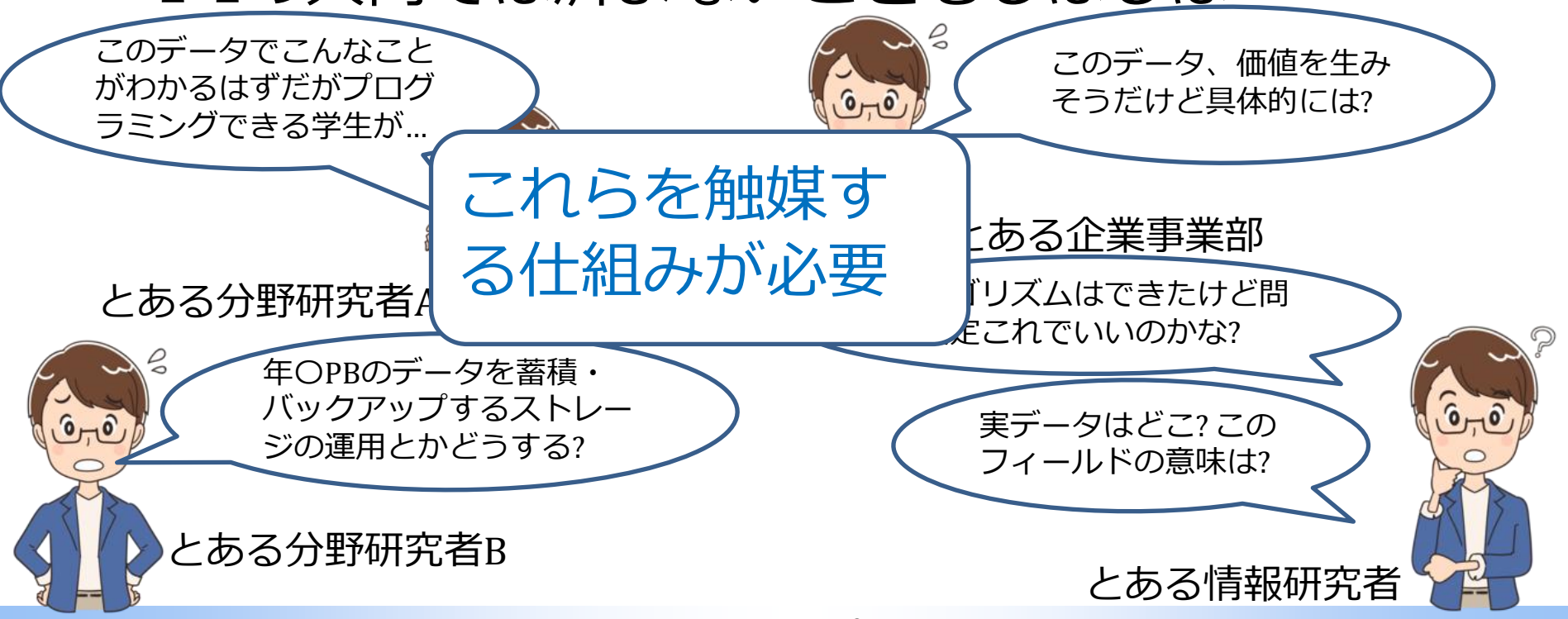
- **シミュレーション+AI（計算手法）**



必要なこと

- データ科学・活用は常に分野やセクタをまたがる横断的な活動
- 1-1の共同では済まないこともしばしば



以降の話

- データ活用社会創成プラットフォーム
- mdx
- システムとしての特徴
- 利用・共同研究・産官学連携の進め方
- パイロットプログラム

Society5.0を実現するためのデータ活用による知識集約型社会の創成
－データ活用社会創成プラットフォームの構築－

データ活用社会における現状認識

- ICT機器の爆発的な普及や、AI、ビッグデータ、IoT等の社会実装が進むなど競争が激化。一部の企業や国のデータの囲い込みにより経済社会システムの健全な発展が阻害される懸念。
- 我が国が成長していくためには、デジタル新時代において、データを我が国全体の共同資産として、スピード感をもったデータ利活用環境の整備が急務。
- Society5.0が目指すインクルーシブな社会を実現するためには、地域における知識集約の中核を担う大学を起点としてイノベーションの創出を図り、知識集約型社会を構築することが重要。
- サイバー空間とフィジカル空間が融合するデジタル新時代において、我が国に蓄積された農業、医療・健康の分野、教育データを含む多くの有用なビッグデータを共同で活用する上で、人材と技術を有する全国の大学を超高速・高信頼で網目状につなぐ国際的優位性をもつSINETを最大限活用することが重要。
- 異種データや異種知識の融合・活用を促進するための「場」として、様々な分野のデータ保持者、解析者、利用者が参画するコミュニティを形成するとともに、データ活用を目指す利用者へのコンサルティングやアプリケーション開発支援が不可欠。

文部科学省における取組

- 経済財政運営と改革の基本方針2018や未来投資戦略2018において重要性が指摘されているリアルデータの利活用を念頭に、データ活用社会創成プラットフォームを推進するため、SINETを通じて収集されるリアルデータの集積や、解析結果を速やかにフィードバックする機能を備えたシステムを整備（2019年度予算）
- 文部科学省と大学コミュニティ、地域社会等が一体的に連携し、全国の国立大学等をハブとしたデータ活用社会創成プラットフォームの実現促進に向けた検討を行うための「データ活用社会創成プラットフォームの推進に関する有識者会合」を設置。
- 地域・産業・社会基盤を支える拠点となる大学を中心として、民間への利用拡大も視野に我が国全体の知識集約型社会の実現に向けた環境「データ活用社会創成プラットフォーム」を構築

データの高度利活用環境（NII・東大に先行して整備）

【設備整備】


IoT接続（モバイル）
AI特化スパコン
リアルタイム処理対応スパコン
高速/セキュアストレージ　等

SINETを通じて、全国のデータ収集・通信・解析環境をオンデマンドで活用。
高度・多様なデータ利活用により新たな価値を創出。

データ活用社会創成プラットフォームの推進に関する有識者会合

リアルタイム処理対応基盤社会創成プラットフォームの実現に向けた実務的な検討を行う場

【主な検討課題】
- リアルタイムデータの解析・活用を目的とした基盤ソフトウェアの研究開発や技術の実証のための基盤システムの整備のあり方
- 産学連携体制（コミュニティ）の構築・強化、その中核としての大学の役割等一体的な連携を確保する仕組み　等

利活用ニーズを踏まえたシステム整備・ソフトウェア開発

【大学等におけるデータ利活用の潜在的なニーズ】
- 地域農業・漁業・観光業のスマート化
- 認知症・生活習慣病などの早期発見、予防方法の提案
- スポーツ科学への応用
- 初中段階から高等教育、社会人教育に至る一貫した教育データの利用　等

大学等連携コンソーシアム

大学を中核としたデータ活用実務機関が連合したコンソーシアム

【主な取組】
- データプラットフォームの活用促進、データ活用ニーズ調査
- コミュニティ間連携の強化・促進等

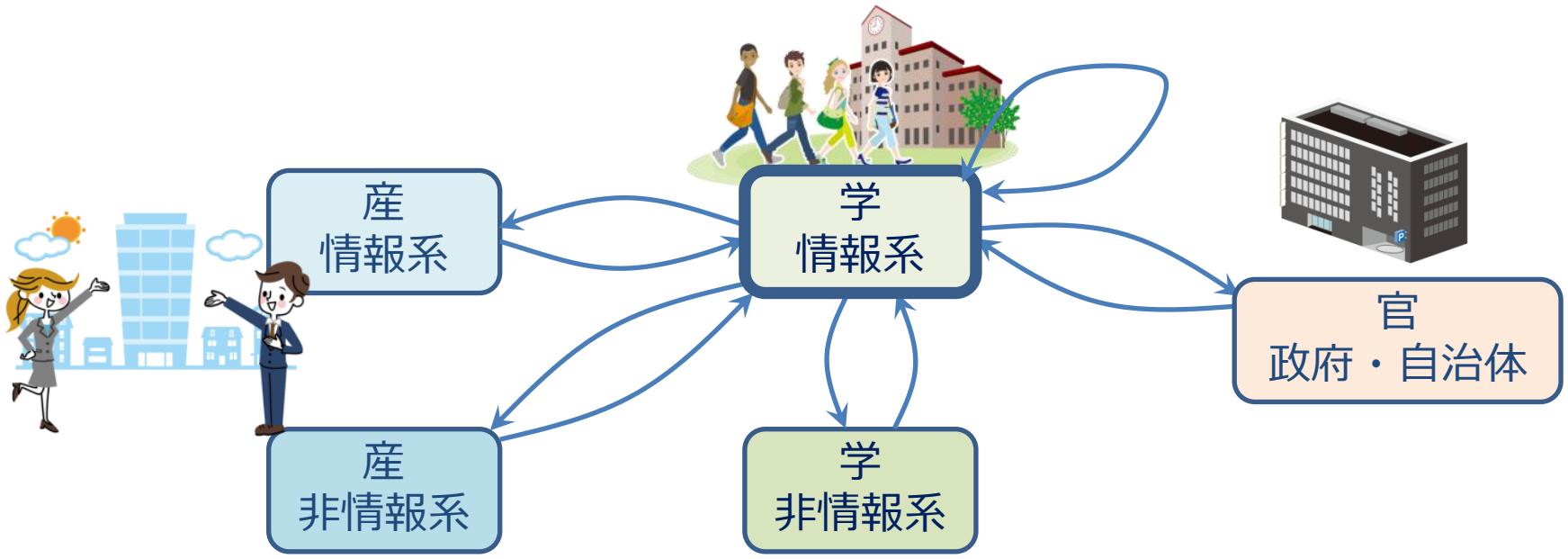
21

データ活用社会創成プラットフォーム

- データ科学・データ活用の｛研究、産官学連携、社会実装｝を進めるための取り組み
- 有識者会合による方向付け
- 2研究所（NII,AIST） + 8大学（北大、東北大、東大、東工大、名古屋大、京大、阪大、九大）で離陸
  - 離陸後、より広い連携体制を構築

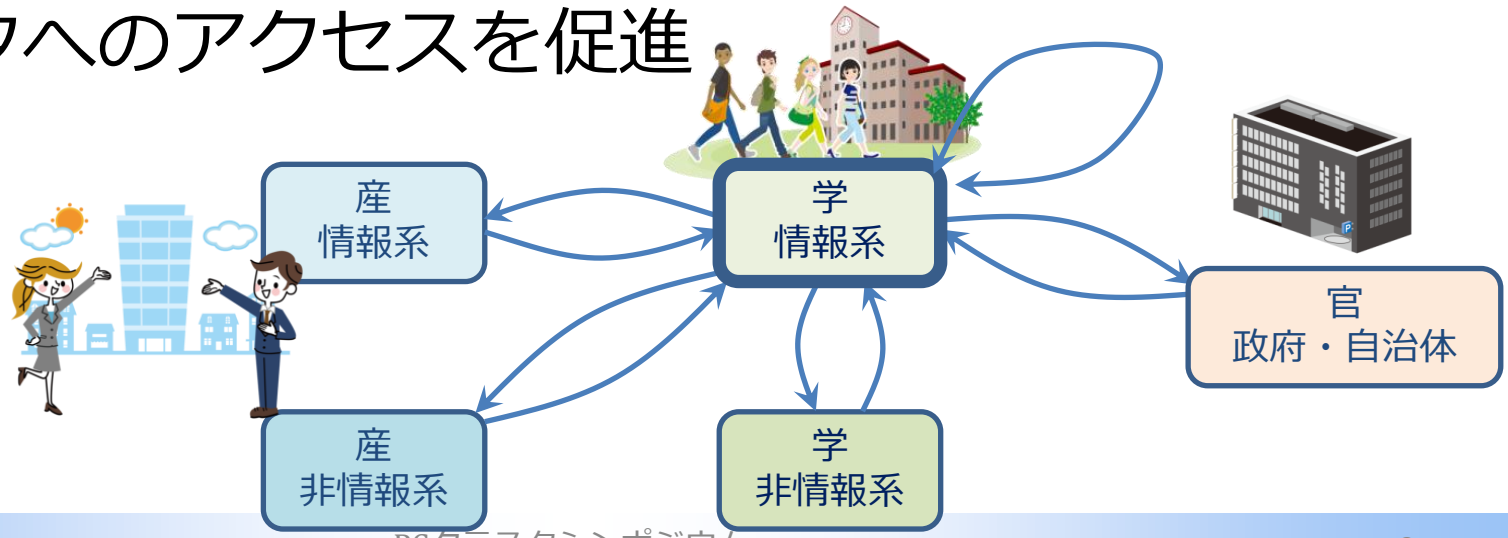
コミュニティ形成・発展

- プラットフォーム＝コミュニティ
  - マシンやデータレポジトリ（だけ）のことではない！
- データ科学・活用での、分野・セクタを横断した連携を触媒するハブとなることを目指す



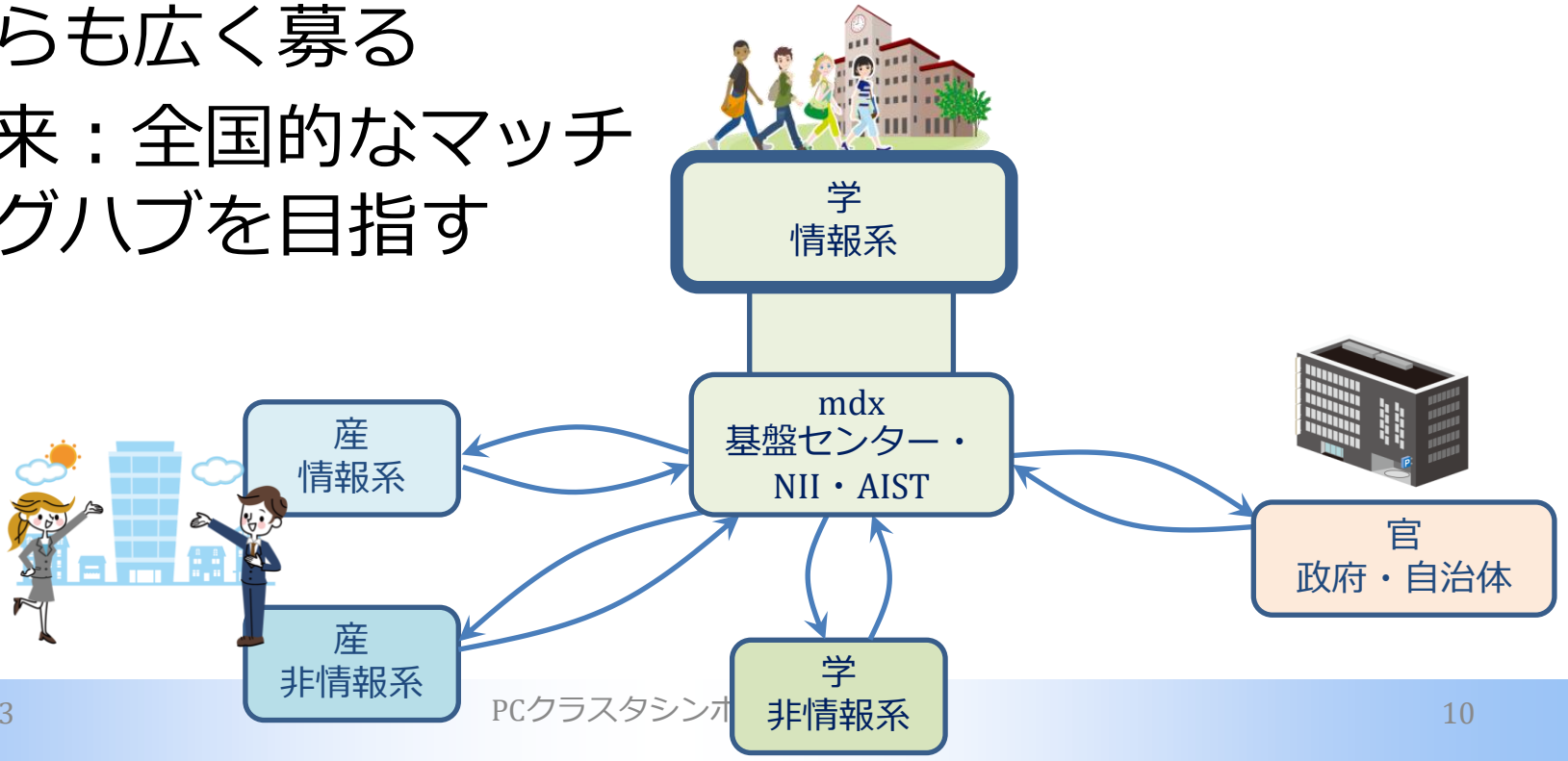
「情報系アカデミア」の役割

- ヒト：情報系専門（特に若い研究者・学生）の知と腕を提供
- ハコ：研究開発・非営利目的への廉価な計算・ストレージ資源を提供
- データ：巨大パブリックデータや利用目的限定データへのアクセスを促進



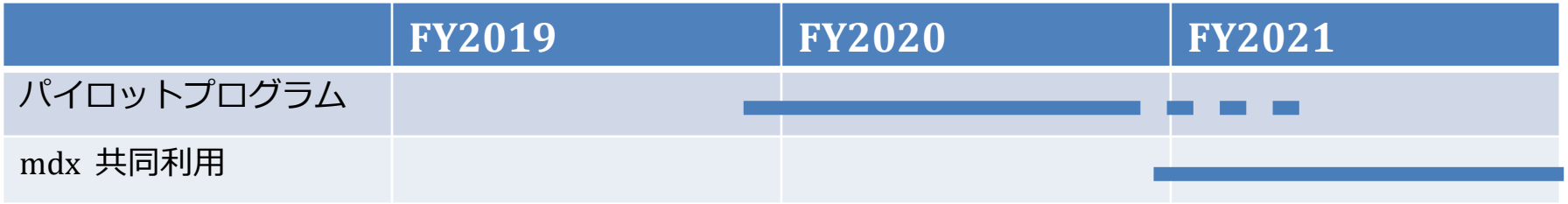
mdx 参画機関の役割

- mdx参画機関 = 基盤センター＋NII＋AIST
  ⊂データ科学系アカデミア⊂情報系アカデミア
- 情報系研究者の参加（ヒト）をmdx参画機関外からも広く募る
- 将来：全国的なマッチングハブを目指す



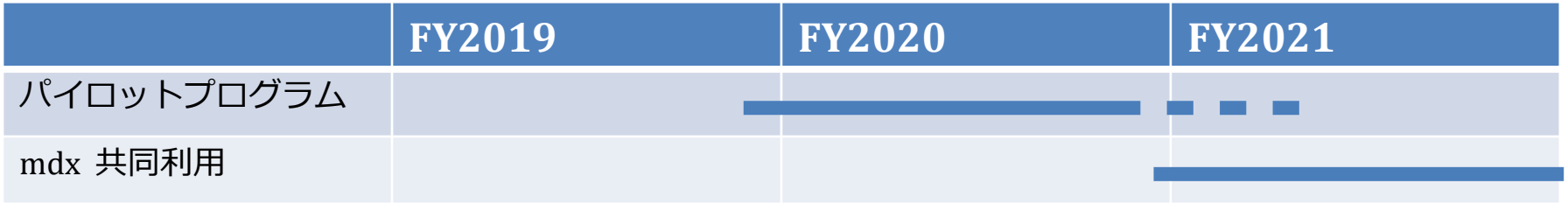
取り組み内容

- データプラットフォーム mdx の構築と運用（運用 2020年度末～）
- 2研究所+8大学一体となって、利用・共同研究・産学連携をオープンに募集
- 運用開始以前の連携も募集（パイロットプロジェクト. 2019年度末～）



取り組み内容

- データプラットフォーム mdx の構築と運用（運用 2020年度末～）
- 2研究所+8大学一体となって、利用・共同研究・産学連携をオープンに募集
- 運用開始以前の連携も募集（パイロットプロジェクト. 2019年度末～）



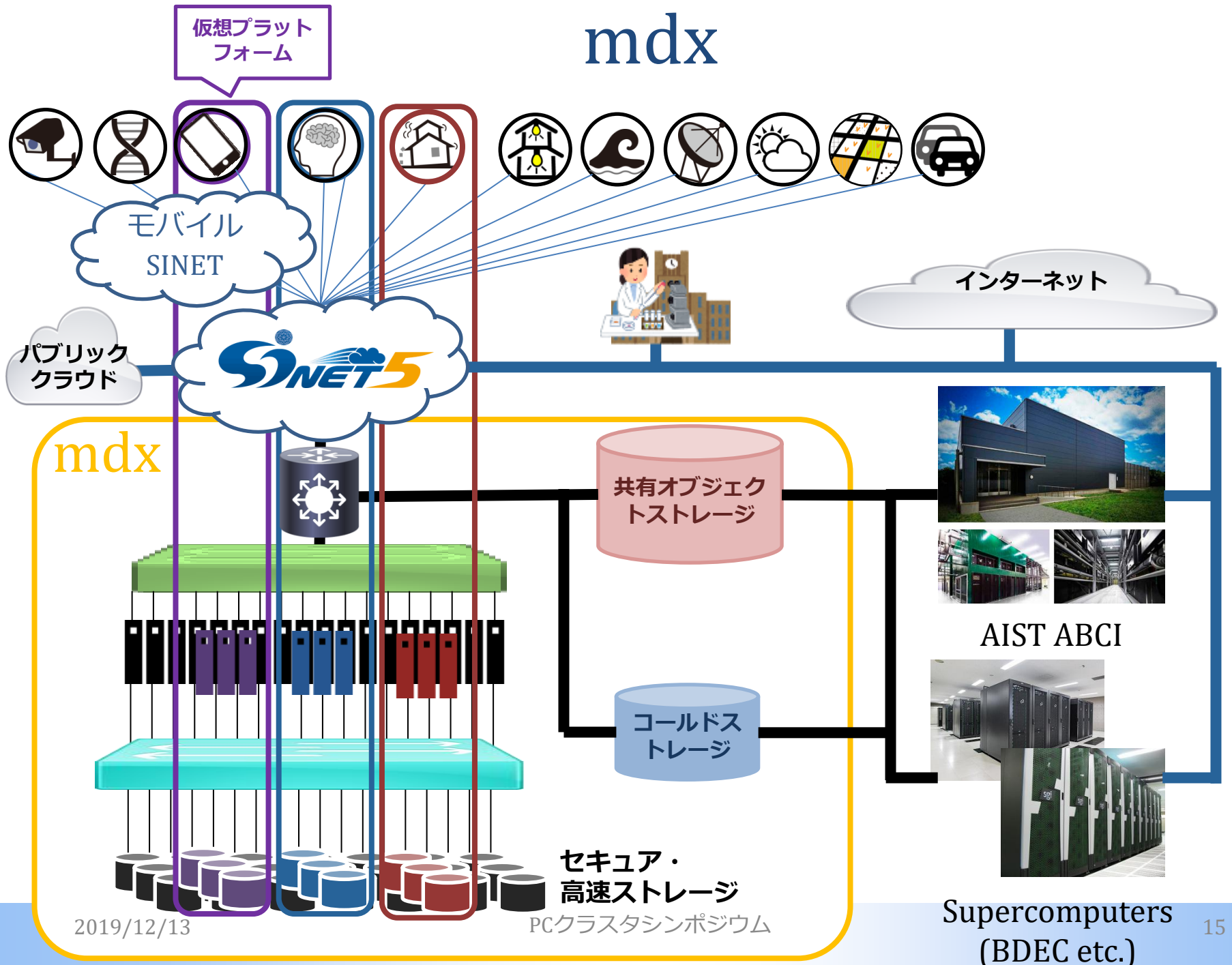
mdx : 環境設計上の鍵

- データ科学・活用のための基盤
- これまでのHPC用途のクラスタとは異なる
- 「単一OS環境 + バッチスケジューラ」では済まない
  - 分野データプラットフォームのホスティング（連続稼働）
  - 多様なソフトウェア構成の許容
  - 長年にわたるデータの蓄積・利用
  - 高いデータセキュリティ・隔離への要求
  - IoT・センサなど外部データのストリーム（実時間）処理
  - JupyterLabなど対話的・探索的利用（AI, 機械学習）
  - データ検索・発見のための利用
  - 対話的利用から高性能環境へシームレスな移行・連携

mdx

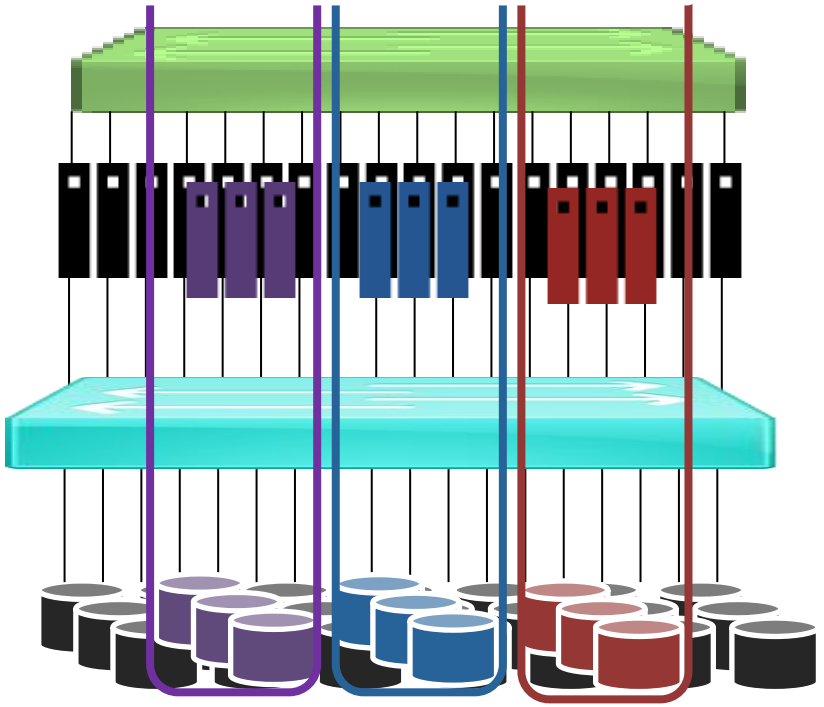
- 仮想プラットフォーム
  - 柔軟・セキュアな環境の構築が可能
- SINET・モバイルSINETと接続
  - セキュアIoT環境の構築が可能
- 高性能計算環境
  - mdx, ABCI, BDEC

mdx


仮想プラットフォーム
モバイル
SINET
パブリッククラウド
SINET5
インターネット
mdx
共有オブジェクトストレージ
コールドストレージ
セキュア・
高速ストレージ
AIST ABCI
Supercomputers
(BDEC etc.)

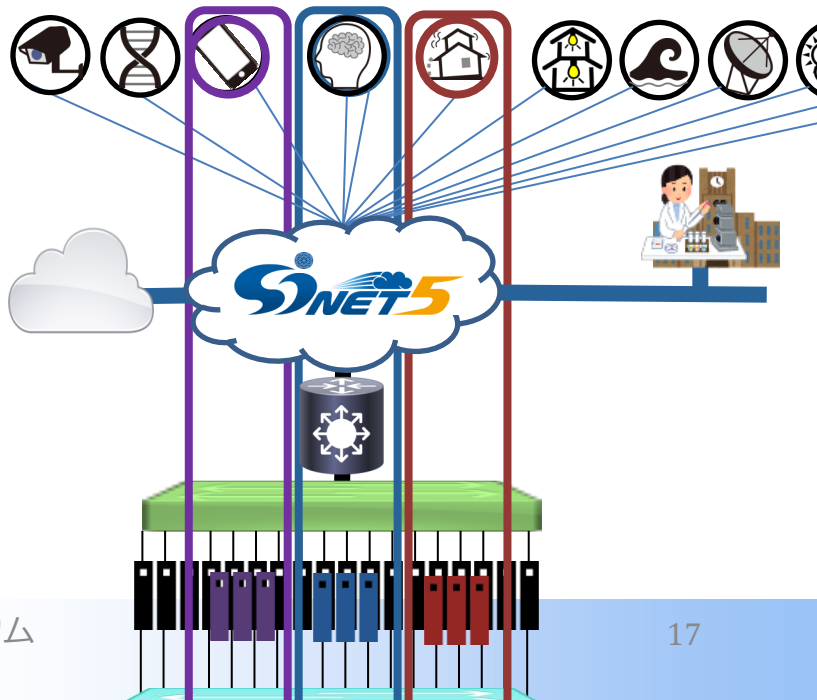
仮想プラットフォーム

- 仮想マシンとVPNを用いて互いに隔離された「疑似占有環境」
- 柔軟性
  - 各プラットフォームごとに自由に（管理者権限で）環境設定可能
  - 常時稼働が必要なデータ公開サービスなどを運用可能
- セキュリティ
  - ひとつの仮想プラットフォームが侵入を受けても他へ影響しない



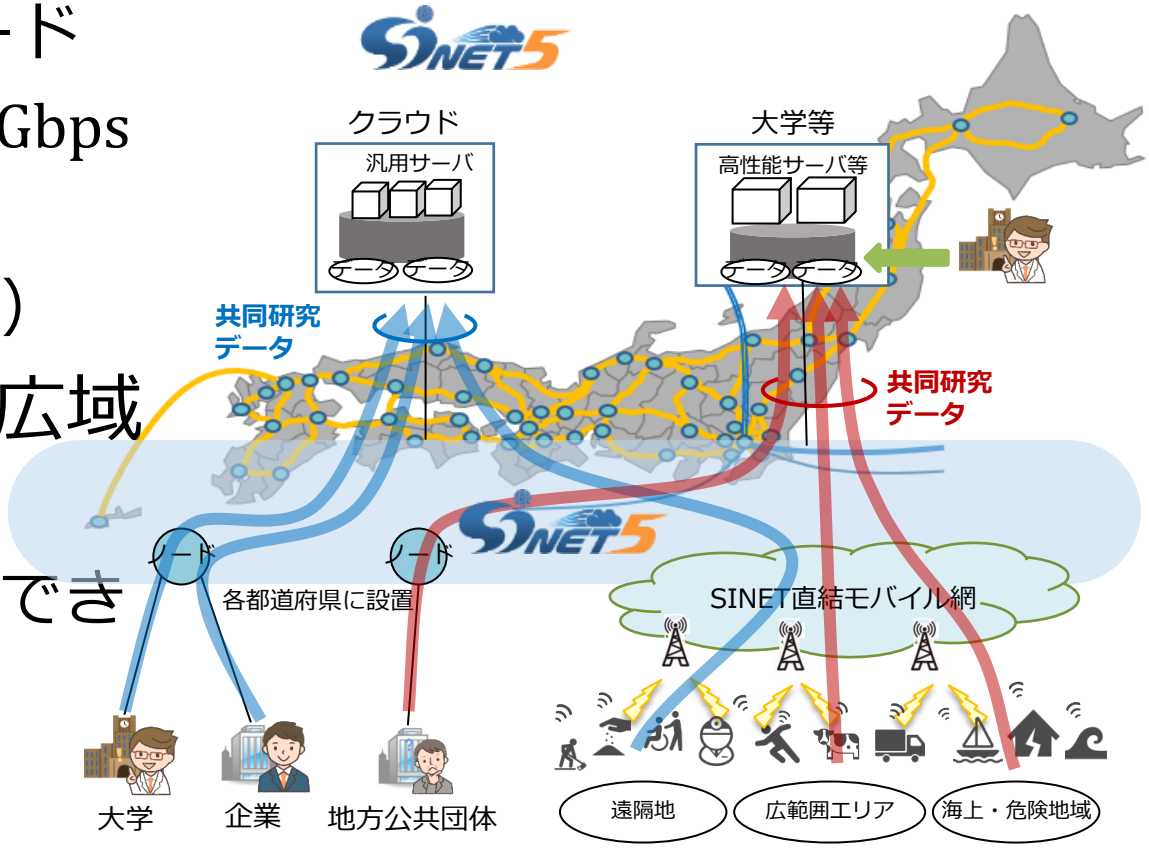
SINET・モバイルSINETとの接続

- 仮想プラットフォーム用のVPNをSINETへ延伸可
- 他のSINETサイト（大学・研究機関）やモバイルSINETとセキュアに接続可能
- とくにIoTデバイスのセキュリティ確保に有用



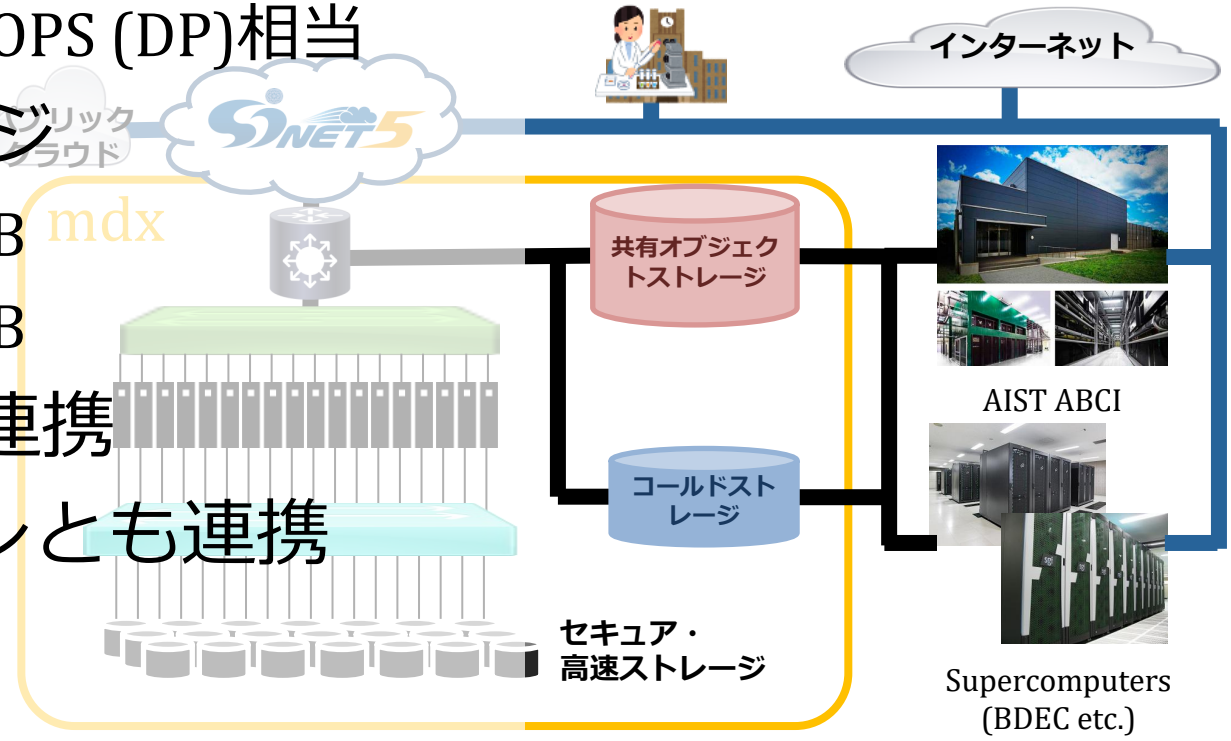
SINET 5・モバイルSINET

- SINET 5（有線）
  - 全都道府県にノード
  - ノード間は≥100Gbps
  - 冗長経路
  - 広域VPN （L3/L2）
- モバイルSINET（広域データ収集基盤）
  - SINET VPNに直結できるモバイル環境



高性能計算環境

- mdx 計算ノード
  - 汎用、学習、推論用ノード
  - 合計 ～ 10 PFLOPS (DP)相当
- mdx ストレージ
  - 内部 ～ 10-15 PB
  - 共有 ～ 15-20 PB
- 産総研ABCIと連携
- 大学のスパコンとも連携



未確定情報を含みます

データ整備と利用促進

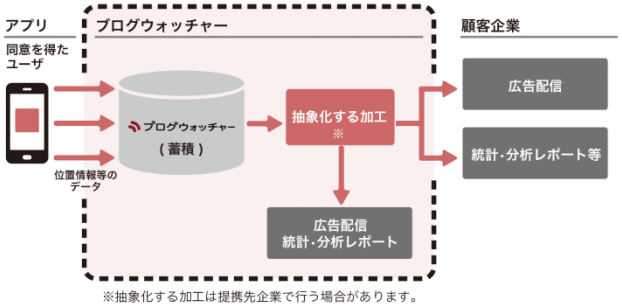
- 研究資源としてのデータを mdx 上に整備、または mdx 上で利用可能とする
  - ダウンロード困難な巨大なパブリックデータ
    - e.g. Common Crawl, MLPerf
  - 連携協力企業提供のデータ
    - mdxとして契約。個々のユーザの利用を簡便化
    - cf. NII情報学研究データレポジトリ
  - 他のデータプラットフォームプロジェクト提供のデータ
  - 強いセキュリティを要求するデータ

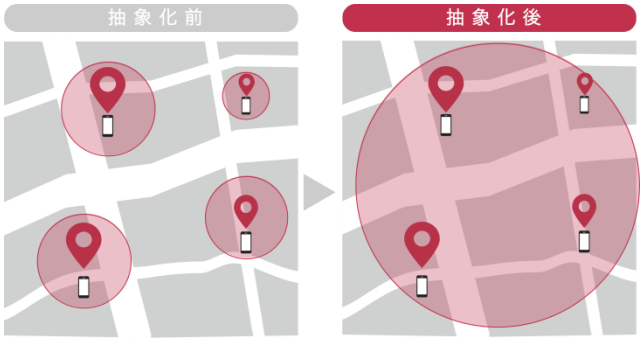
ブログウォッチャー

ブログウォッチャー位置情報データ

- ユーザ同意に基づきスマートフォンアプリから取得した（匿名、抽象化後の）位置情報
- 研究目的での利用促進、新しい利活用の共創、プライバシーの心配の少ないデータ・モデリングの研究、etc.

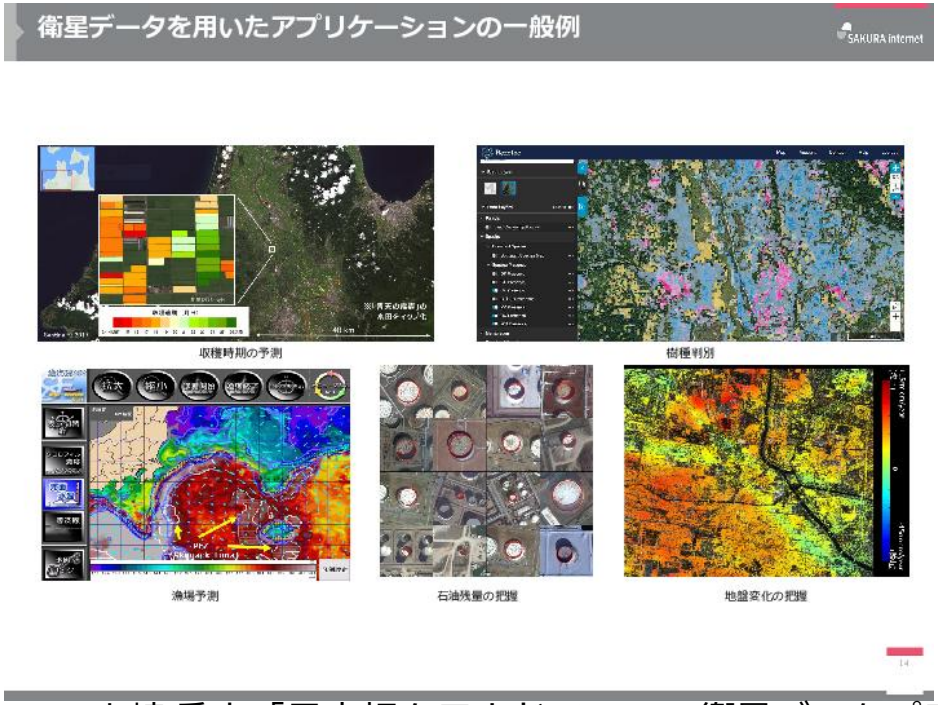
Profile Passport DMP

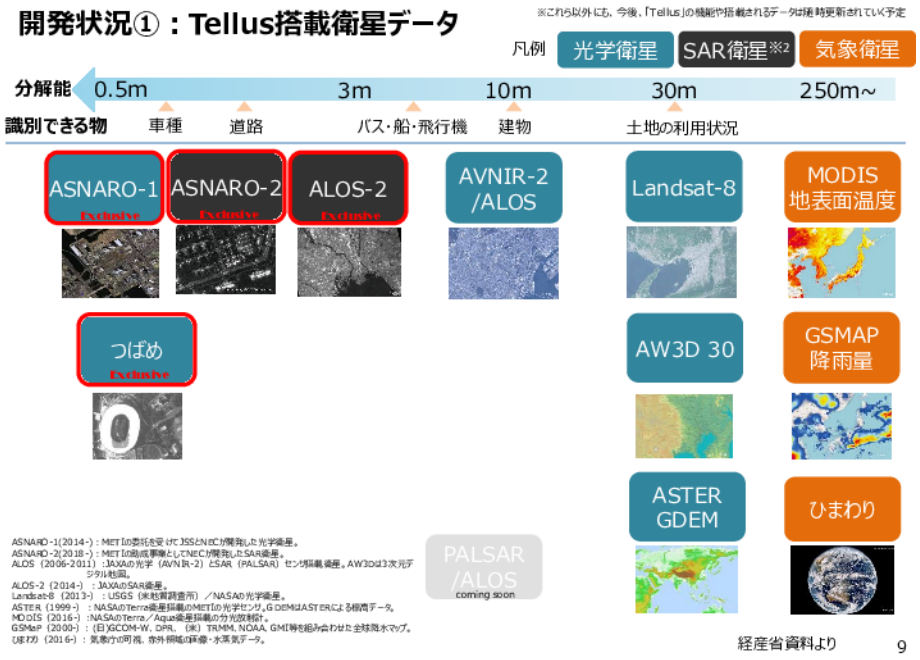




Tellus衛星データ

- 衛星データプラットフォームTellusのデータと連携
- 研究目的での利用を促進、新しい利活用共創

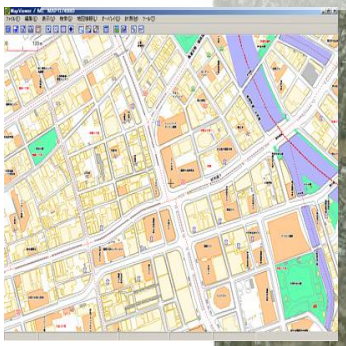




山崎 秀人「日本初クラウドベースの衛星データプラットフォーム「Tellus」」
https://cloudconference.jaipa.or.jp/app/download/16943630896/specialty_session.pdf?t=1568343337

全国スケール・超高詳細の航空画像とデジタル地図

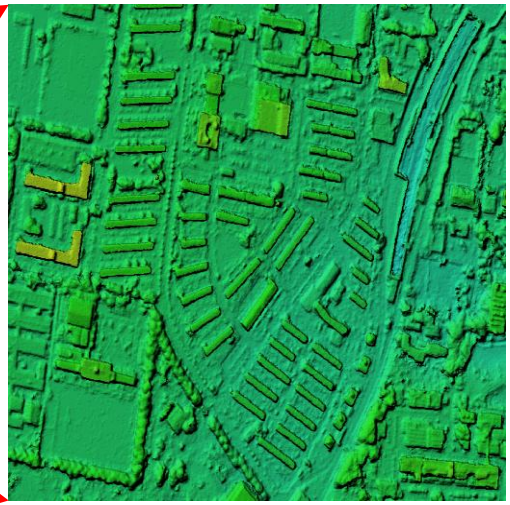
- 東京大学空間情報科学研究センター
- 超高詳細15cm/pixel, ほぼ毎年更新
- 不動産や都市解析など多面的な利用に期待



デジタル地図

15cm/pixelの航空画像（モザイク済み）

航空画像から3次元計測された地形・建物形状

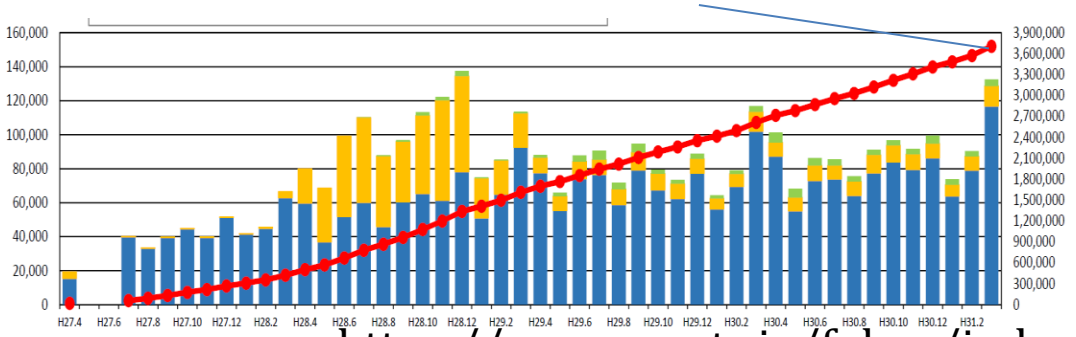


未確定情報を含みます

ETC 2.0プローブデータ

- ETC 2.0
  - 高速道路上（1600か所）に設置されたITSスポットと車載器が双方向通信
  - （匿名）各車両の経路（断片）情報が利用可能

ETC 2.0車載器台数 400万台@2019年2月

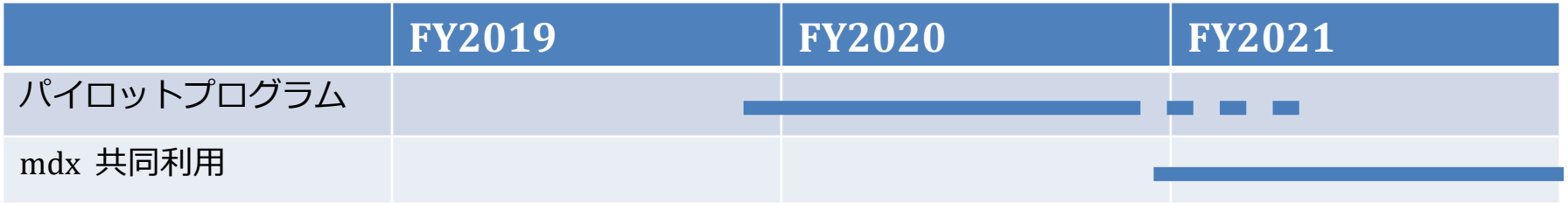


https://www.go-etc.jp/fukyu/index.html

未確定情報を含みます

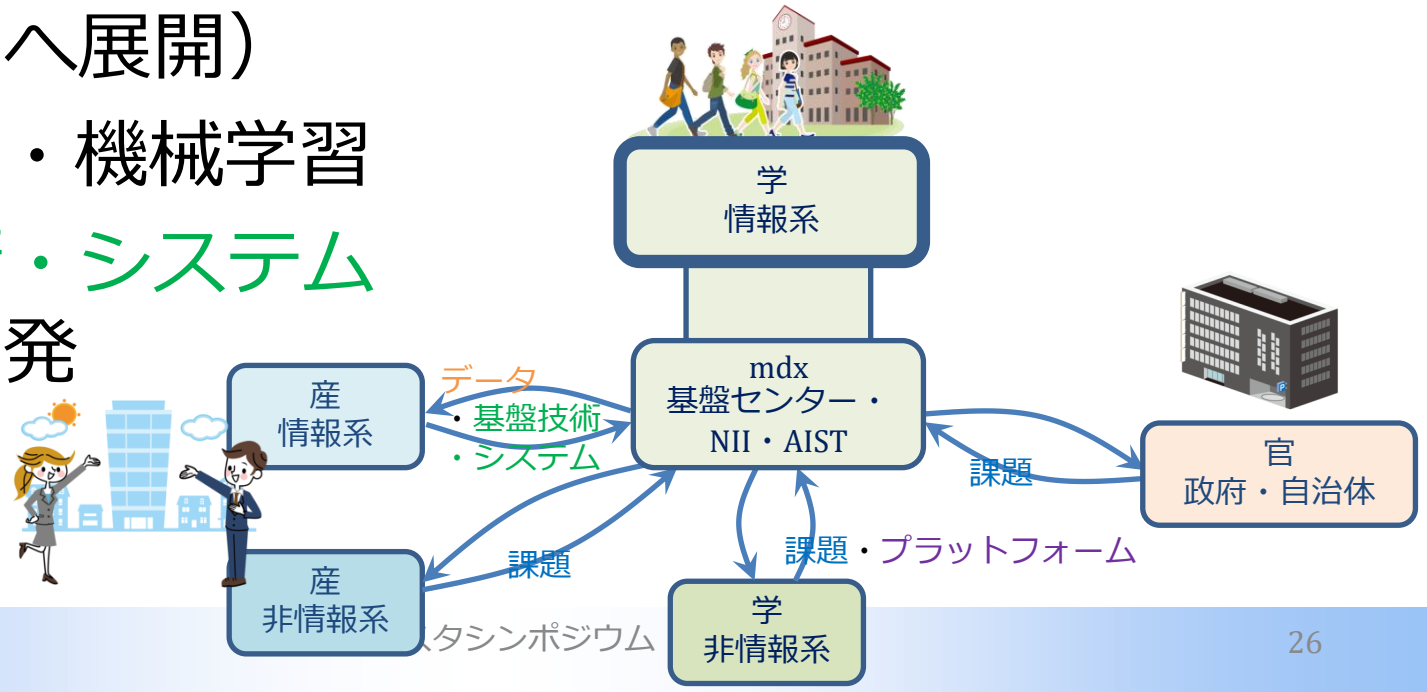
取り組み内容

- データプラットフォーム mdx の構築と運用（運用 2020年度末～）
- 2研究所+8大学一体となって、利用・共同研究・産学連携をオープンに募集
- 運用開始以前の連携も募集（パイロットプロジェクト. 2019年度末～）



多様な利用・連携を歓迎・追及

- データ活用課題持ち込み（分野の、地方の、企業の、…）
- 分野データの整備・プラットフォーム構築
- データ提供・サービス（収益）化模索（⇒ 商用クラウドへ展開）
- 高性能AI・機械学習
- 基盤技術・システム研究・開発
- …



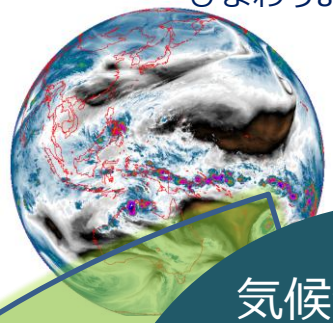
水と気候の大規模データ研究拠点

東京大学ビジョン2020：[研究] 1-①国際的に卓越した研究拠点の拡充・創設
[教育] 2-④付置研究所等の教育機能の活用
[社会貢献]3-①　学術成果の社会への還元

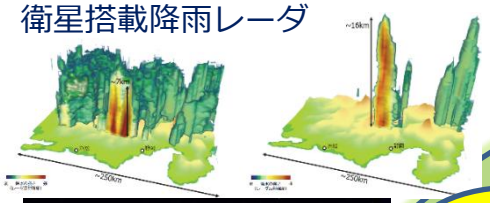
多様な高精度地球観測衛星

雲と降水の特徴を決定するプロセスの理解

ひまわり8号



衛星搭載降雨レーダ



プロセス解明

気候変動と水循環

極端降水・干ばつをもたらす雲降水現象の解明

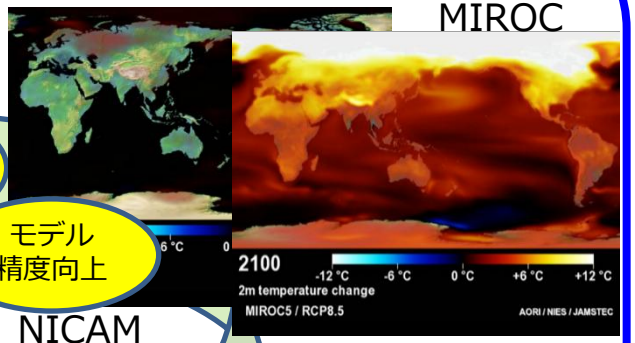
大気海洋研が開発する世界最先端気候モデル

予測

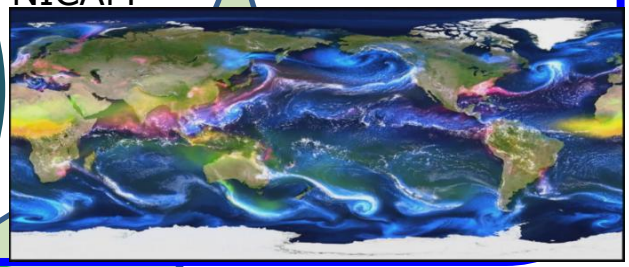
同化

モデル精度向上

MIROC



NICAM



AI

SINET

学内外の連携

外部機関の大規模データ

衛星データ	気象データ	数値モデル
ひまわり8号 GPM GCOM-W ….	世界の再解析 アメダス 高層データ	数値予報 気候予測 アンサンブル実験 CMIP, d4PDF

ICT利用の機関連携による大規模データ解析

- 分散型データアーカイブ体制の構築
- 大規模データ解析
- 人工知能（AI)の利用

プロジェクトの３つの柱

- 大規模データハンドリング体制の構築
- 雲降水プロセス・極端降水の研究
- 人材育成・社会貢献

情報基盤センターの（これまでの）オープンな共同研究方式

- 8大学の情報基盤センター群は共同利用・共同研究拠点 JHPCN を運営（2010～）
- オープンな共同研究を募集・支援してきた実績

JHPCN

JHPCNの共同研究の成立過程

応募書類

JHPCN

拠点

- マッチングは応募前になされている
- 似た分野・知った研究者同士の共同向け

データ科学・活用のための「課題持ち込み」型のマッチング

- 課題を持つユーザの分野も多様化・拡大
- 一言でいえばこれまでより「情報系との距離が大」な分野｛との共同・への支援｝が重要
  - データと研究課題・活用アイデアを持っている
  - 「MPI + OpenMP !? 知らない、要らない」場合も
  - 機械学習・大規模データ処理・ストリームデータ処理の知見やプログラミング・ツール支援が必要
- 人文・社会・経済・医療・薬学・・・

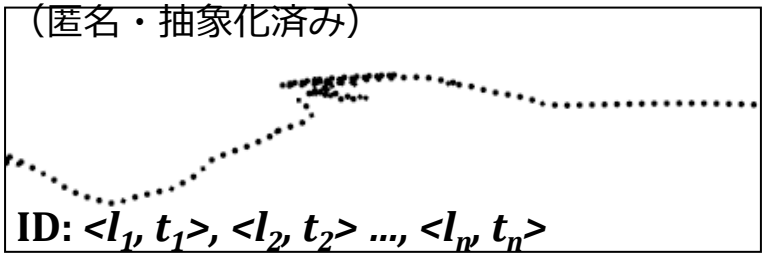
GPSデータと感染データ連携

- 東大大学院薬学系研究科ITヘルスケア社会連携講座
- 人の移動がインフルエンザの感染拡大にどう寄与しているかをモデリング、流行予測モデル構築へ

位置情報データ

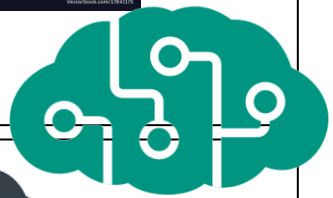
BlogWatcher

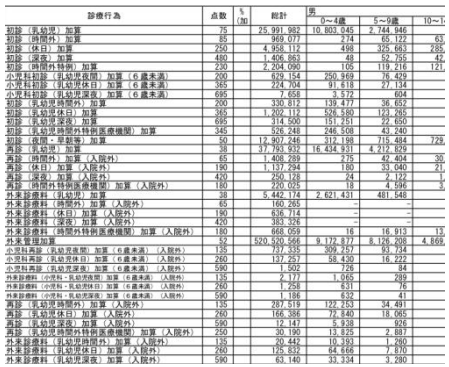
「提携アプリをダウンロードし、位置情報の取得を許可したユーザー」のスマートフォン端末から、GPSで補足した位置情報（匿名・抽象化済み）



ID: $<l_1, t_1>, <l_2, t_2> ..., <l_n, t_n>$

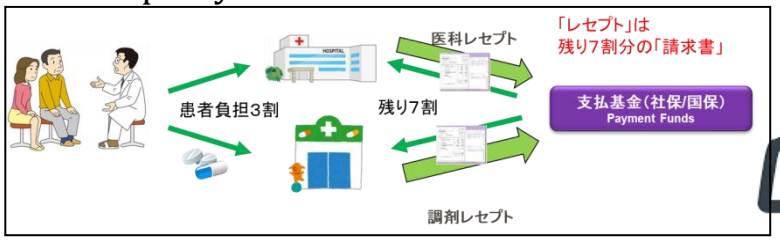
空間データ・データモデリング専門家





NDB, JMDC

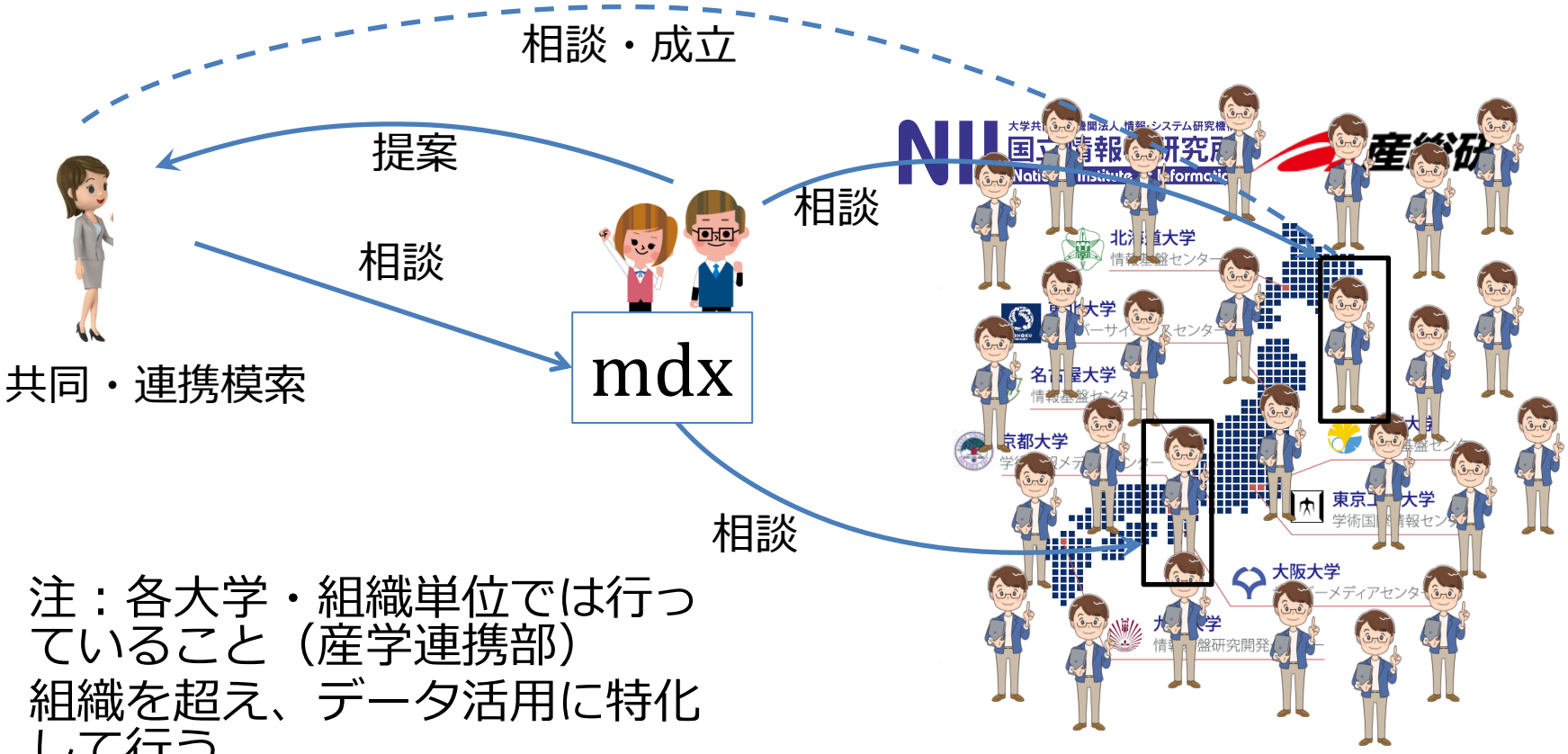
レセプト情報・薬の処方情報などインフルエンザ感染数のproxyとなるデータ



医療データ　**インフルエンザ感染・流行情報**　**医療データ・感染学専門家**

mdxのマッチング

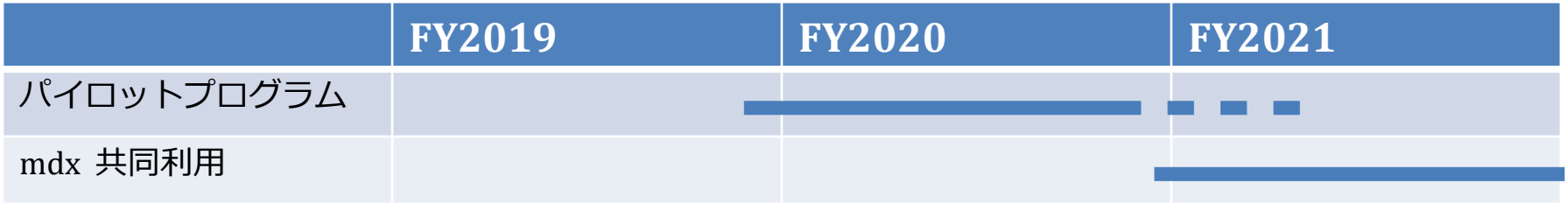
- チーム成立前のマッチングの相談をmdxとして受付



- 注：各大学・組織単位では行っていること（産学連携部）
- 組織を超え、データ活用に特化して行う

取り組み内容

- データプラットフォーム mdx の構築と運用（運用 2020年度末～）
- 2研究所+8大学一体となって、利用・共同研究・産学連携をオープンに募集
- 運用開始以前の連携も募集（パイロットプロジェクト. 2019年度末～）



パイロットプロジェクト

- mdx稼働（2020年度末）時に展開される活動のうち、システム稼働前に行える部分の前倒し実施を支援
  - 必要な経費
  - 既存の計算資源
  - 共同研究者のマッチング
- 年内アナウンスを予定

まとめ

- NII, AISTと8大学（北大、東北大、東大、東工大、名大、京大、阪大、九大）は共同で、
  - データ科学・活用のための基盤 mdx を導入
  - 共同研究・産官学連携の仕組みを運用
- mdx ≈
  - 仮想プラットフォーム
  - VPNでの隔離
  - セキュアIoT（モバイルSINET）
  - 高性能計算機・ストレージ
- 共同研究・連携募集 ≈
  - 課題持ち込み
  - 分野プラットフォーム構築
  - データ提供
  - 基盤技術研究
  - etc. など広く募集

多くの方々（情報系・非情報系・産・官・学）の参加・協力をあおぎながら進めていきます
データを持つ人
活用する人
システムを作る人