mdx:データ活用社会創成プラットフォームの紹介

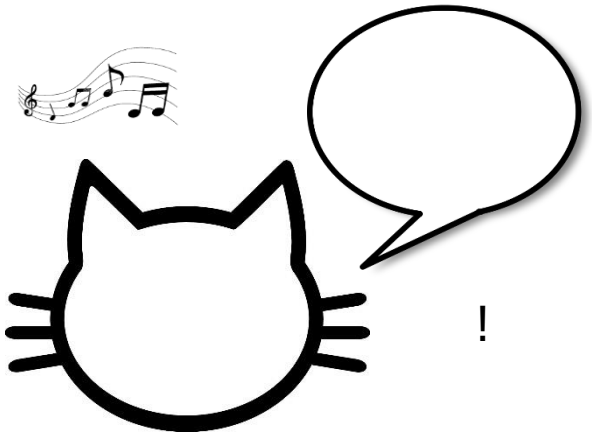
小林博樹
東京大学

自己紹介

- 現職 東京大学
  - （本務） 情報基盤センター データ科学研究部門 部門長
  - （兼任） 空間情報科学研究センター（本務 ~2020/3）

- 専門分野
  - 情報デザイン、センサネットワーク（動物向け）、サウンドスケープ

- 趣味
  - 環境音を聴いて、動物の表情を見て、（面白い）吹き出しを考えること。



東京大学情報基盤センター

- 4研究部門
  - 情報メディア教育
  - データ科学
  - ネットワーク
  - スーパーコンピューティング
- 全学・全国ミッション
  - キャンパスネットワーク
  - セキュリティ
  - 高性能計算機
  - 教育用環境

全国の大学の情報基盤（的）センターについて

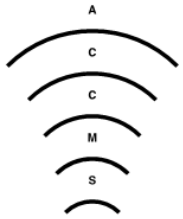
- 多くのセンターが
  1. 情報（主に基盤的）分野の研究
  2. 学内利用の情報基盤（ネットワーク、教育など）
     - 例えばオンライン授業のためのIT導入...
  3. 全国利用の情報基盤（スパコンなど）、それをハブにした学際研究（情報ｘ〇〇学）

を三軸とする点で共通

- 3.のため国の「共同利用・共同研究拠点」という仕組みで8大学のセンターが共同して拠点を運営している

8大学が連携して運営する共同利用・共同研究拠点（JHPCN）について

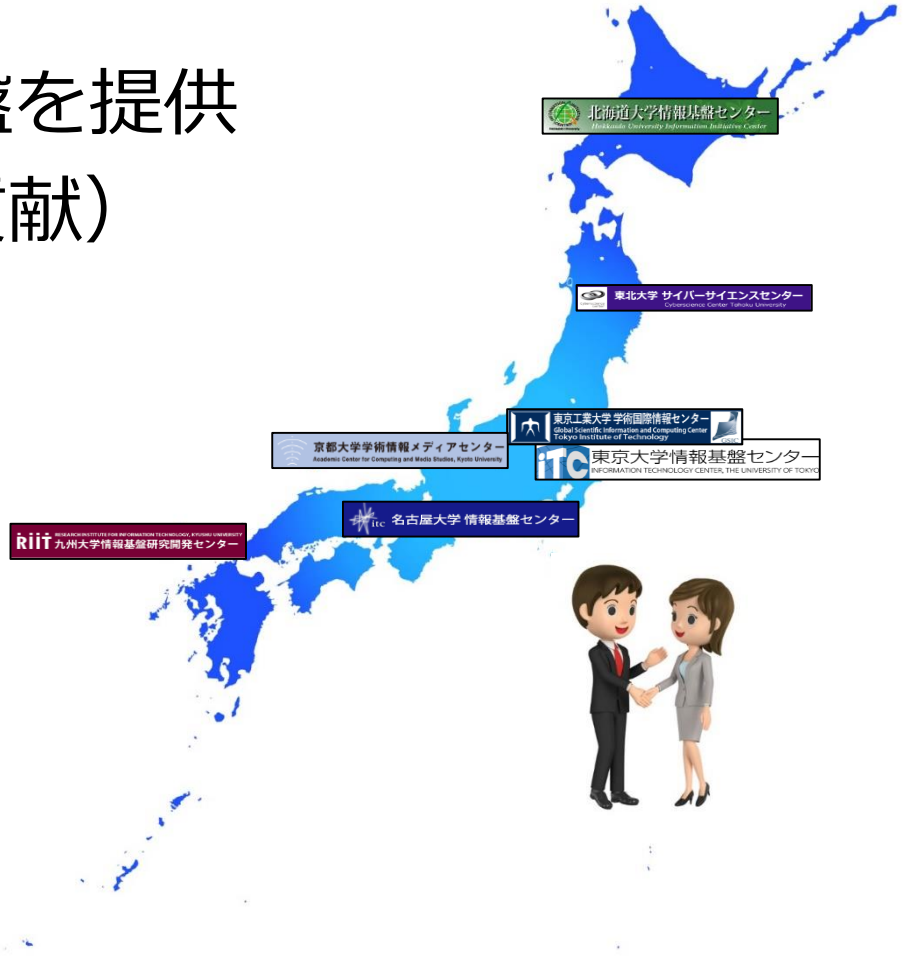
- 8大学の基盤センターで共同運営する共同研究の基盤
  - 北大、東北大、東大、東工大、名大、京大、阪大、九大
- 全国から共同研究を募集
  - 採択された課題に計算設備利用時間（主にスパコン利用時間）を割り当て
  - 現在シミュレーション中心の計算科学分野が多く集まる
  - 企業利用も受け付けています



基盤センターと拠点の使命

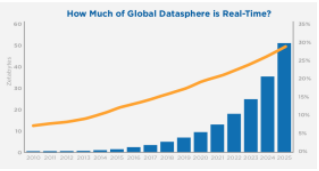
- 情報の専門家と情報基盤を提供
- 情報学 x ○○学（学際貢献）
- コミュニティ形成

今日の話はこの思想をこれまでのシミュレーション中心・計算科学から、分野も基盤もセクターも「データ中心」に広げるための**構想の紹介**



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」）**

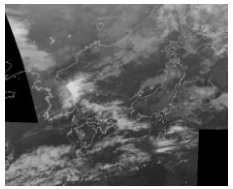
PyTorch　Chainer

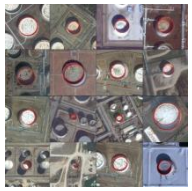
jupyter　docker

kaggle　aws

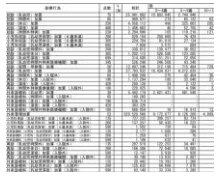
- **社会応用に直結するデータの整備**


GPS


気象


衛星

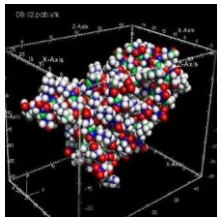

公共交通

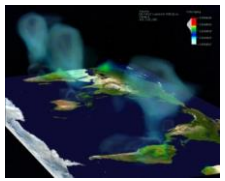

医療行為

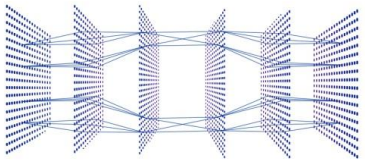

道路・インフラ

- **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）**

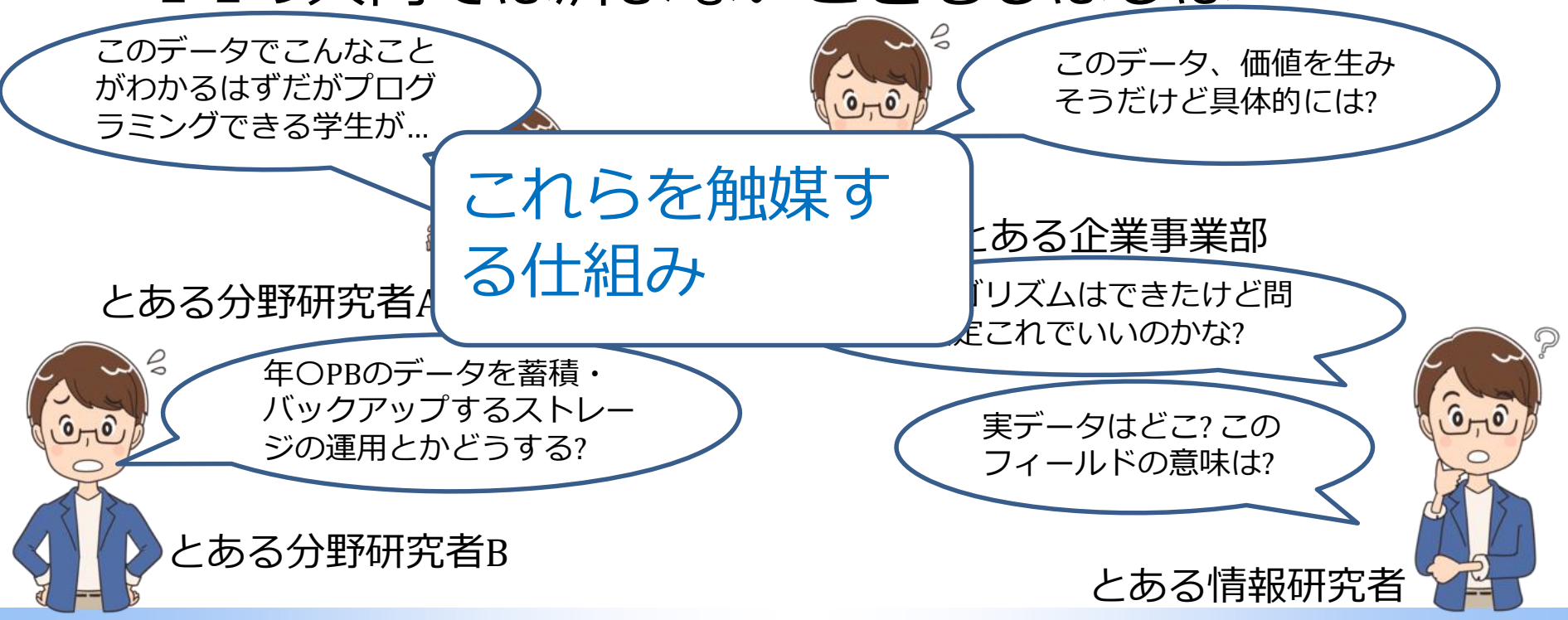






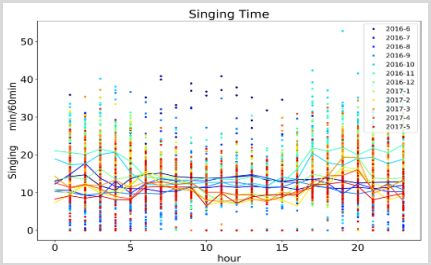
必要なこと

- 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



研究事例の例：電源・情報・道路インフラが存在しない未除染・高線量地帯に生息する野生動物IoTを用いた環境・健康情報センシング（帰還困難区域内）

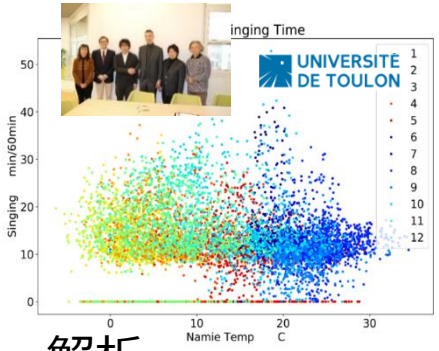
http://www.town.namie.fukushima.jp/soshiki/2/namie-factsheet.html

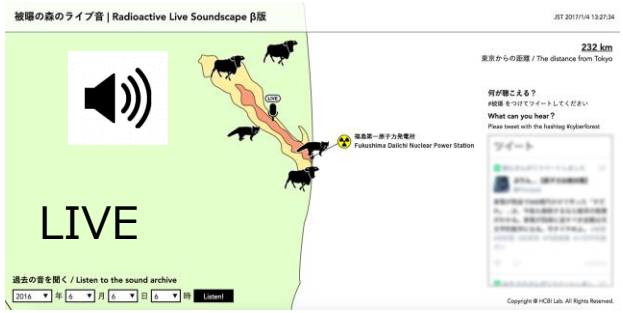


被曝動物の健康情報（鳴き声）の変化の可視化9000時間分

プロジェクト用に整備されたシンクノード基盤・電柱・通信回線
NII　SINET広域データ収集基盤 実証実験

野生動物Iotの評価実験中の様子





今この瞬間の帰還困難区域の音が聴こえるWEBサイト

収集・通信・解析

モバイル　　ビックデータ　　リアルタイム（＋安定性）

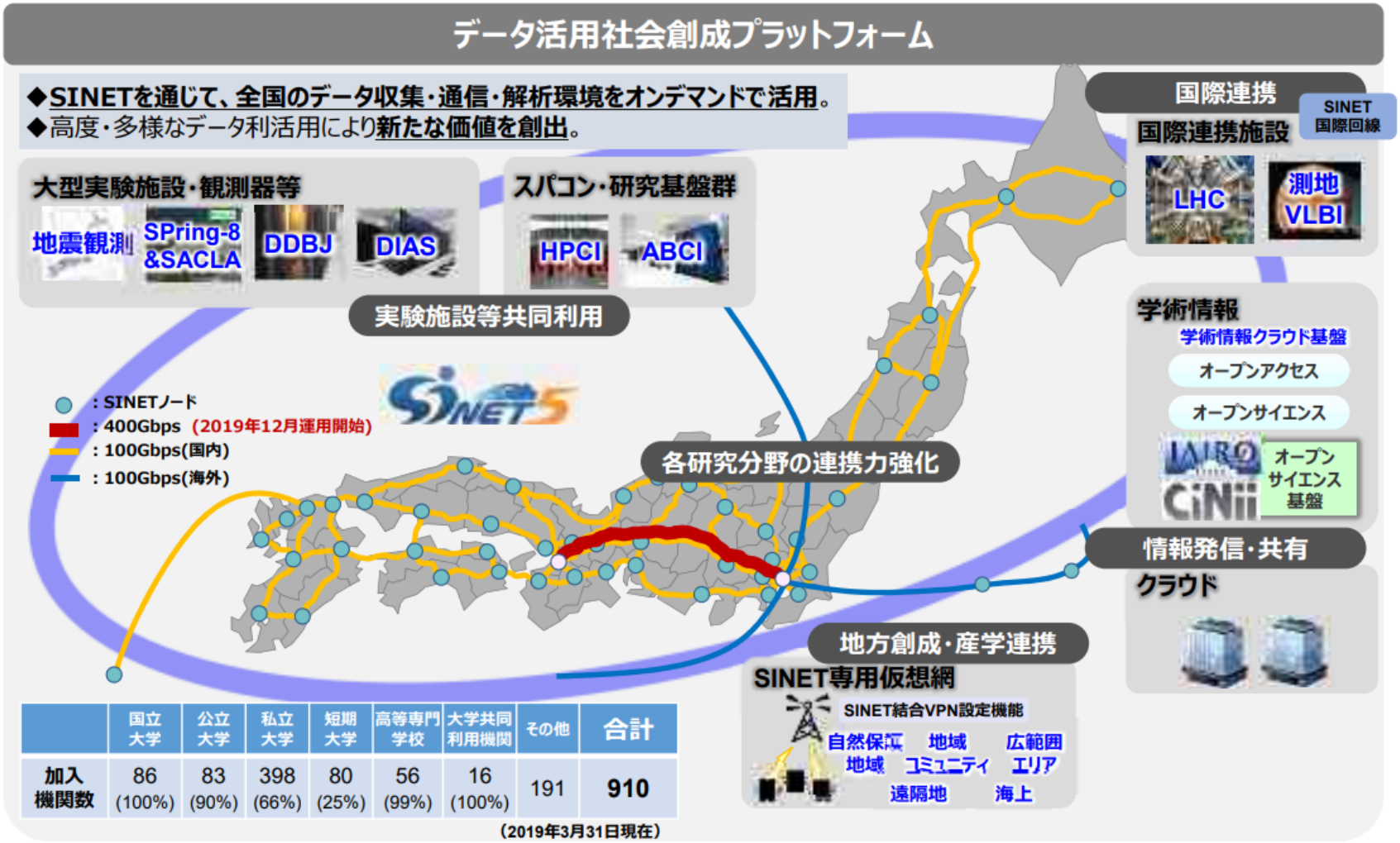
以降の話

- データ活用社会創成プラットフォーム
- mdx
- mdxの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

（政策的文脈）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） + 9大学（北大、東北大、筑波大、東大、東工大、名古屋大、京大、阪大、九大）で離陸
- JHPCNが中心的役割を果たすが、それ以上にコミュニティ（ユーザ・基盤側とも）を広げる

大学・研究機関等の情報基盤インフラ整備・活用例



	国立大学	公立大学	私立大学	短期大学	高等専門学校	大学共同利用機関	その他	合計
加入機関数	86 (100%)	83 (90%)	398 (66%)	80 (25%)	56 (99%)	16 (100%)	191	910

（2019年3月31日現在）

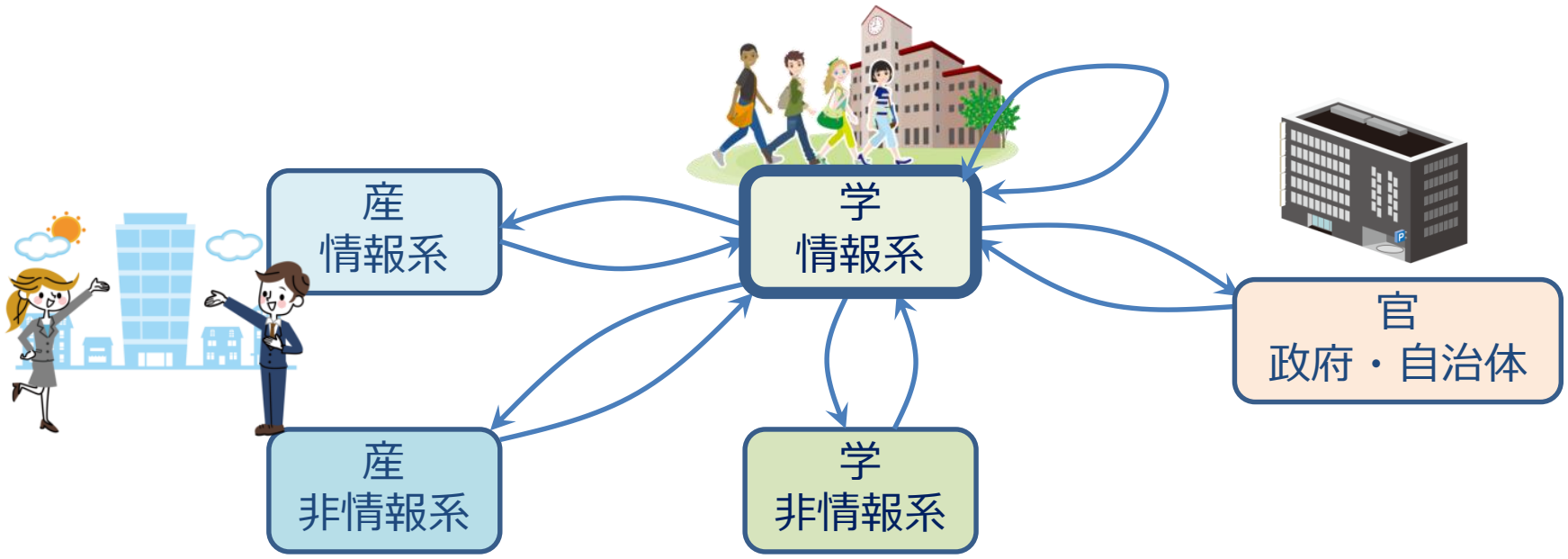
25

https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu29/siryo/__icsFiles/afieldfile/2019/07/12/1418982_6_1_1.pdf

文部科学省:「情報分野における政策の動向について」

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



以降の話

- データ活用社会創成プラットフォーム
- mdx
- mdxの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

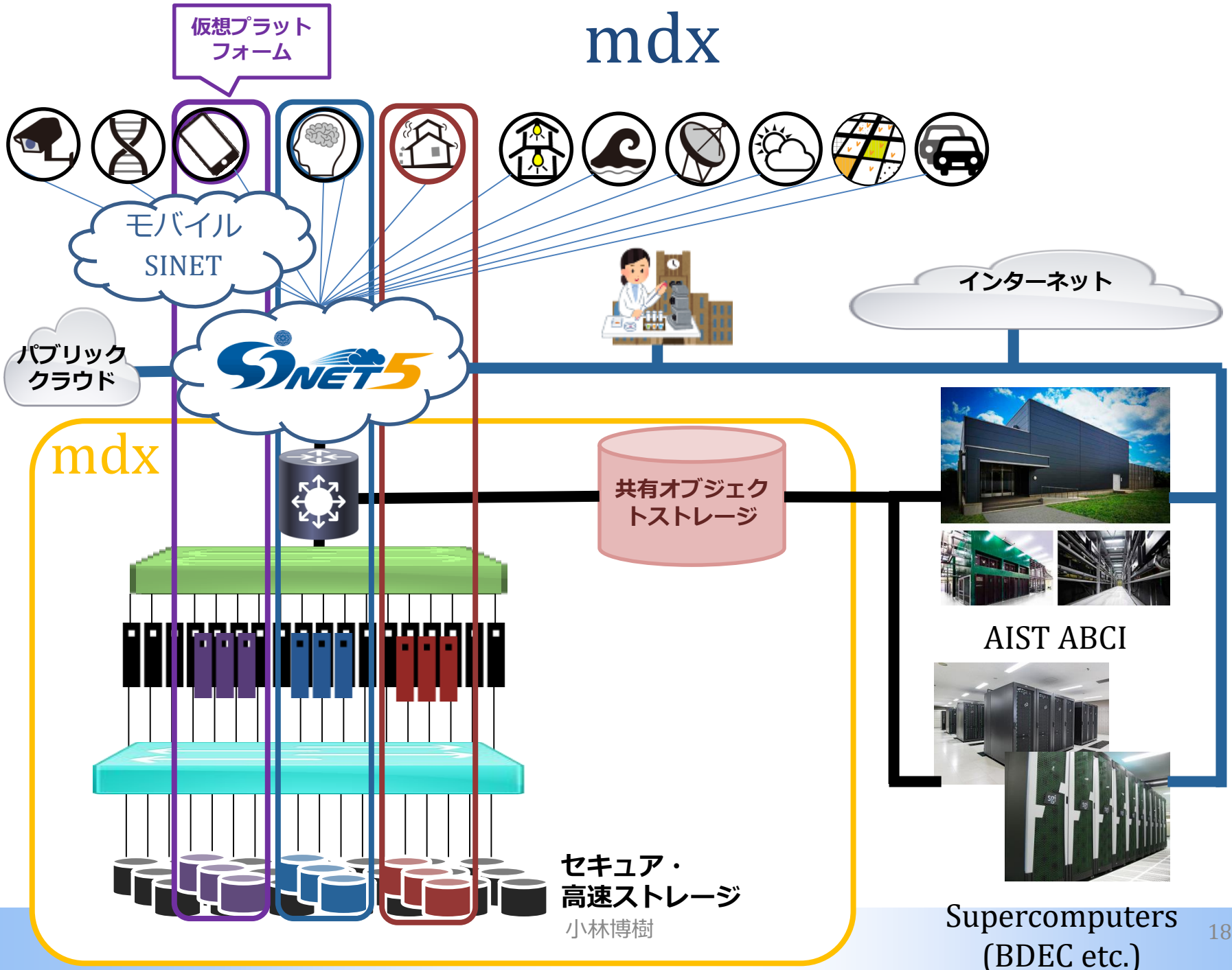
mdx とは
データ{活用・科学}のための基盤

- 「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
- ユーザが標準（常識）と考える環境も異なる
  - 機械学習フレームワーク、コンテナ
  - JupyterLabなどWebベース対話的利用（AI, 機械学習）
  - データ検索・発見のための利用
  - IoT・ストリーム・データ収集など広域・実時間処理
- 多数のデータ・システムの連携が前提（1システムでは完結しない）

mdx 特徴

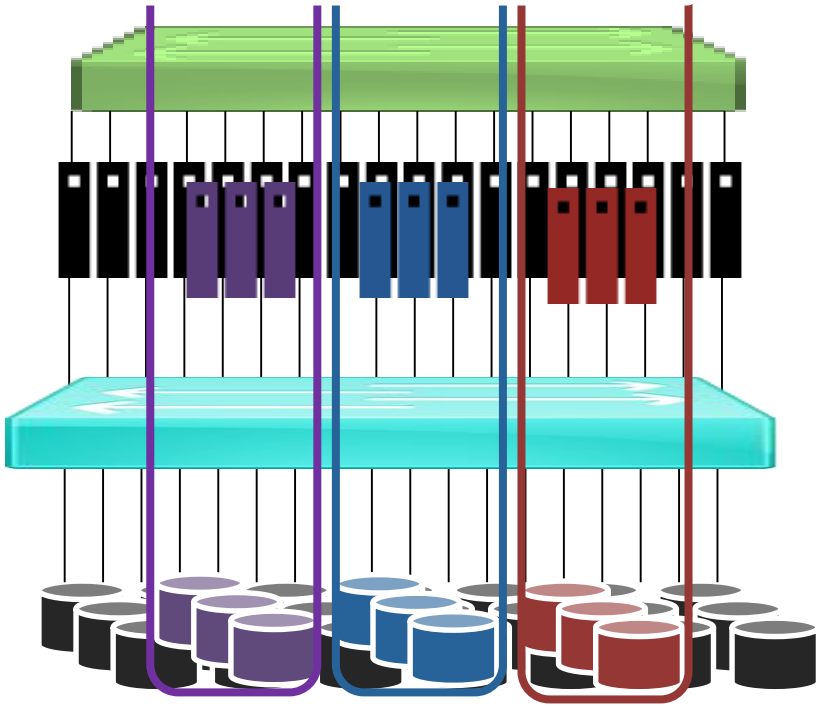

- 2020年度内（末）稼働 @ 東大 柏IIキャンパス
- 仮想化された（複数テナント）環境
- 計算ノード
  - 350+ CPUノード
  - 300+ GPUs
- ストレージ
  - 15+PB （含 NVMe SSD 1PB）ファイルシステム Lustre Multi-tenant拡張
  - 10PB S3オブジェクトストレージ
- 高性能な仮想化されたネットワーク
  - SR-IOV, PVRDMA


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
AIST ABCI
セキュア・
高速ストレージ
小林博樹
Supercomputers
(BDEC etc.)

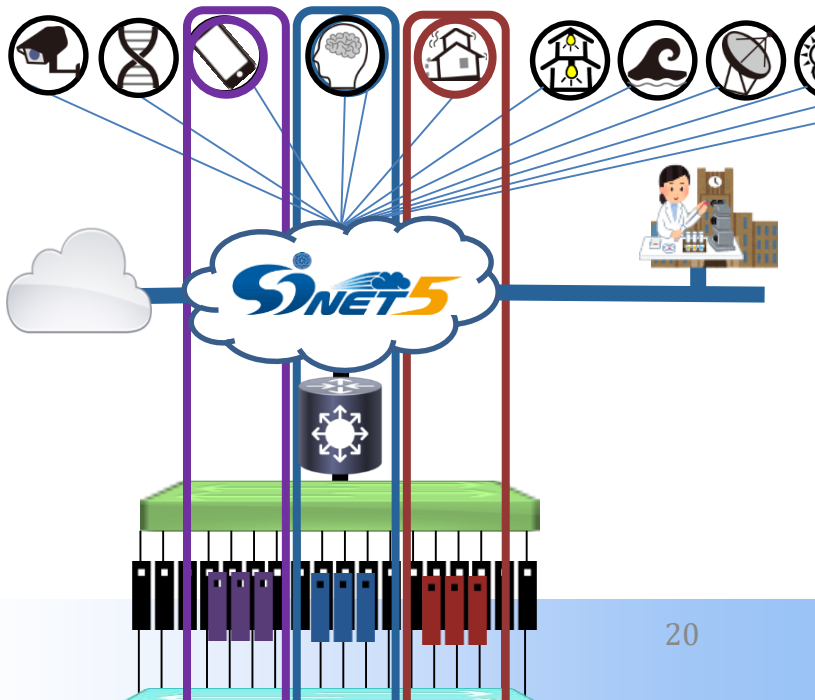
仮想プラットフォーム

- 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



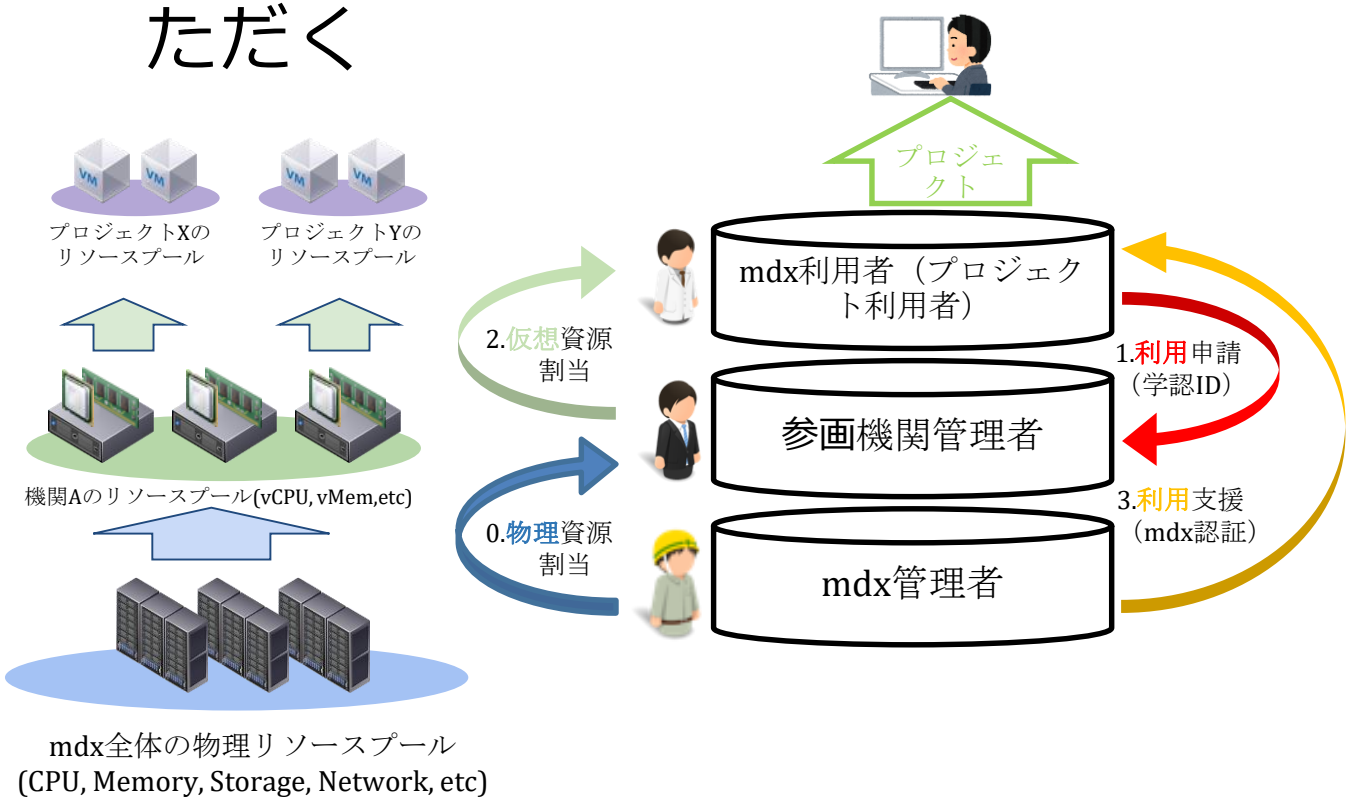
SINET・モバイルSINETとの接続

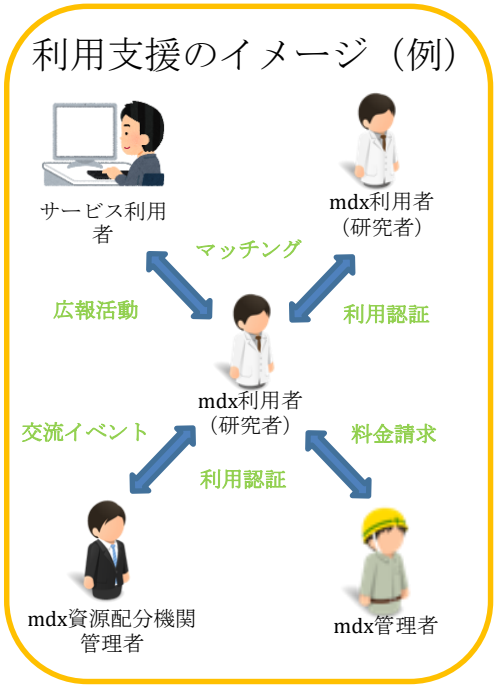
- 仮想プラットフォーム用のVPNをSINETへ延伸可
- 他のSINETサイト（大学・研究機関）やモバイルSINETとセキュアに接続可能
- とくにIoTデバイスのセキュリティ確保に有用



資源の割り当て方

- 各機関（9大学+2研究所）に一定量の資源配分権を割り当て
- 各機関でそれを利用したユーザ獲得に使っていただく





データ連携

- GakuNin 学認
  - 大学をまたがる認証基盤
- GakuNin RDM GakuNin RDM
  - 研究データ管理
  - データのカタログ化、共有・公開設定
  - 外部クラウド・S3ストレージとの連携機能
- mdx 両者を用いて
  - 研究者がサインアップし、すぐに使い始められる環境
  - データの検索から、ブラウザベース処理、高性能処理までをシームレスに実行できる環境

様々な関わりレベル利用者・利用法

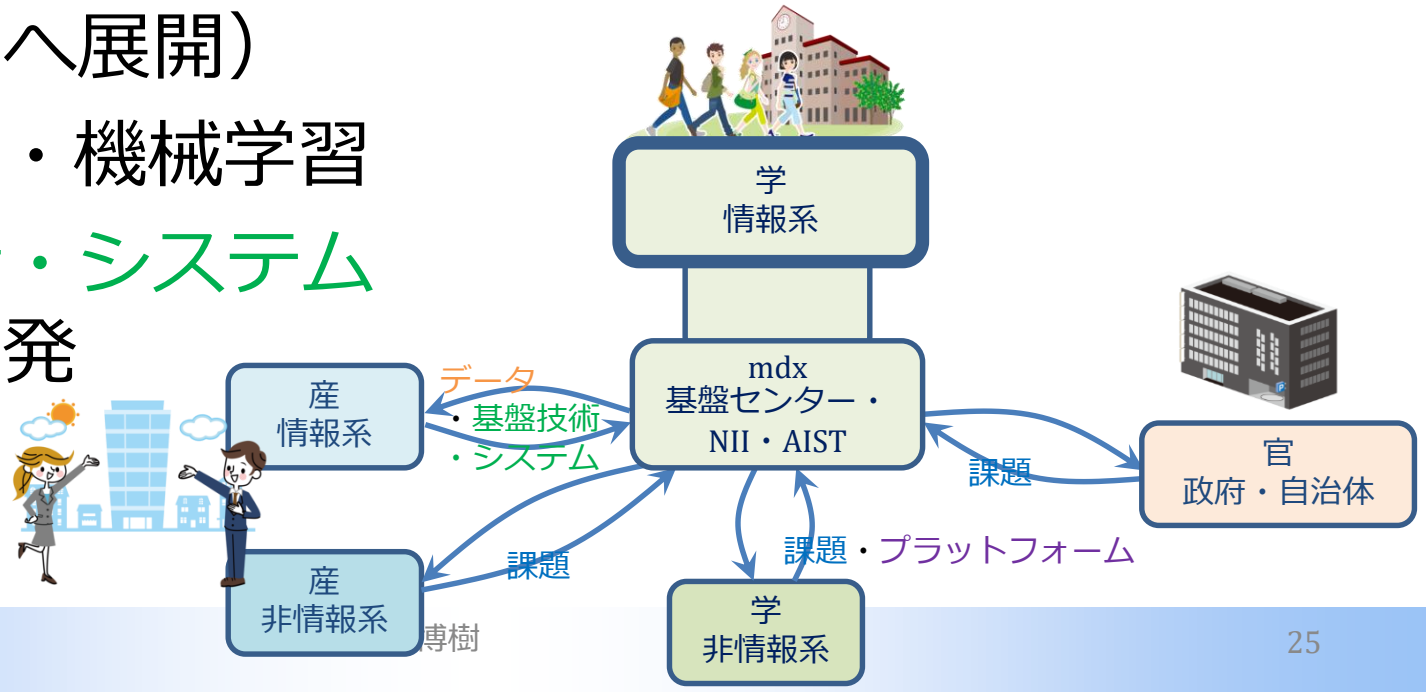
- データ検索レポジトリとしての利用
- Webブラウザベースのお手軽データ解析・AI環境としての利用（cf. Kaggle, Tellus）
- 高性能計算・データ処理環境としての利用
- （主に個人研究者・小規模グループ）中規模データや学習モデル共有ストレージとしての利用
- （主に組織）大規模データ収集&データプラットフォームホスティング環境としての利用（cf. DIAS）
- システム研究のための（仮想化を生かした）自由度の高い環境

利用募集の外形・運用について

- データ検索や少数ノード利用：統一認証基盤などを用い申請書なし・迅速な利用を可能に
- 大規模、本格的な利用：一定の審査
  - ただし新しいユーザ層・分野にリーチする必要性などから当初は現行のJHPCN課題申請ほど厳格な審査は行わない予定
  - 審査基準も現行とは変える必要がある
  - 複数年での利用を可能に
- 戦略的なパートナーシップに基づく利用
  - mdx参画の各機関の裁量を重んじつつ全体で決定・承認
- 当初は（おそらく）課金はしない
  - 並行して将来的な新しい課金モデルを検討
  - 「ストレージの占有量に応じて」というモデルは再考

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
- …



目指すのはセンターマシンの新しい標準形

- データがシステムを横断して検索でき、見つかったデータがどのマシンでもすぐに処理可能
  - 個々のセンター＝マシン＋データ
  - データは（たとえばGakuNin RDMで）見つかる
  - データを特定の人、グループと共有できる
    - ストレージやアクセス方法に依存する（チャレンジ）
  - 見つかったデータがどのセンターマシンでも処理できる
    - SINETの広帯域を生かす
    - ストレージの耐負荷性能、長距離広帯域ネットワーク上のストレージ性能などチャレンジ

まとめ

- 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。mdxはそのための基盤
- NII, AISTと9大学（北大、東北大、筑波大、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

ご清聴ありがとうございました

- 以下は参考情報

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